Sequencing through thick and thin

Sequencing through thick and thin: historiographical and philosophical implications

James W. E. Lowe

Science, Technology and Innovation Studies, University of Edinburgh, UK.

Old Surgeons' Hall, High School Yards, Edinburgh, EH1 1LZ. United Kingdom.

Email: james.lowe@ed.ac.uk

Abstract

DNA sequencing has been characterised by scholars and life scientists as an example of 'big', 'fast' and 'automated' science in biology. This paper argues, however, that these characterisations are a product of a particular interpretation of what sequencing is, what I call 'thin sequencing'. The 'thin sequencing' perspective focuses on the determination of the order of bases in a particular stretch of DNA. Based upon my research on the pig genome mapping and sequencing projects, I provide an alternative 'thick sequencing' perspective, which also includes a number of practices that enable the sequence to travel across and be used in wider communities. If we take sequencing in the thin manner to be an event demarcated by the determination of sequences in automated sequencing machines and computers, this has consequences for the historical analysis of sequencing projects, as it focuses attention on those parts of the work of sequencing that are more centralised, fast (and accelerating) and automated. I argue instead that sequencing can be interpreted as a more open-ended process including activities such as the generation of a minimum tile path or annotation, and detail the historiographical and philosophical consequences of this move.

Highlights:

- DNA sequencing is primarily understood by a 'thin sequencing' perspective.

- I propose a 'thick sequencing' perspective.

- Thick sequencing includes different stages of assembly, evaluation and annotation.

- An alternative picture of the nature and organisation of sequencing is presented.

Keywords:

Genomics; sequencing; physical mapping; DNA; assembly; annotation

1. Introduction

Dominant narratives concerning genomics have hitherto focused on the process and product of determining the order of bases (adenine, thymine, cytosine and guanine; A, T, C and G) along a given DNA strand. I call this the 'thin' sequencing perspective, and contrast it to a 'thick' perspective that encompasses all of the scientifically and technically important processes, procedures, materials and stages leading to intermediary and never-quite-complete sequence products. These thick sequences can potentially be used as a resource by various end-user communities not (necessarily) involved in the practices leading to the production of those sequences. These practices may include improving assemblies of sequences by closing gaps and correcting errors and annotating the sequences to indicate where genes lie on chromosomes. Thick sequencing draws our attention to procedures such as these just as much as the determination of the raw sequence that thin sequencing concentrates on; they are vital in ensuring that sequences can be used more fruitfully as a resource. Thick sequencing therefore calls attention to those processes and methods that do not themselves constitute DNA sequencing, but condition what is sequenced, how sequence data are compiled into assemblies and the augmentation of the sequence data to enable it to relay more information than a long string of bases.

I argue that interpretations of the science of genomics that foreground speed, acceleration, large-scale operations and automation are a product of a thin characterisation of sequencing that primarily concerns the procedures that take place in automated machines and associated computers. In the initial era of whole genome sequencing, sequencing machines were often arranged in parallel in factory-style centralised genome sequencing centres, an approach pioneered by J. Craig Venter at Celera and John Sulston at the Sanger Institute (Sulston and Ferry, 2002; Venter, 2008). This sequencing is therefore characterised by large-scale centralised facilities with automated sequencers, mainly staffed by technical employees. The thin picture of genomics has shifted somewhat, with the lowering of the cost of sequencing per base pair making work in large centralised facilities seem less necessary. This work is additionally devolved (for reasons of convenience or cost) to the corporations that build the machines (e.g. Illumina, Pacific Biosciences) and service-oriented laboratories. The process is nonetheless mainly automated, fast, and organised in an industrial way.

'Thick sequencing' is based on the idea that there is no final product, and that the work and insight required to create any publicly available sequence cannot be fully captured under a thin understanding of sequencing. In this interpretation, sequencing can include creating genome libraries, establishing a detailed physical map, and producing and validating the statistical tools and software required for analysis. As much as determining the base order, thick sequencing encompasses ongoing assembly to increase the size of contiguous stretches of sequence and close gaps, revision, modification, resequencing of particular areas of interest, improving the quality and coverage, verifying and comparing with other sequences. It is about the creation of annotated sequences indicating the position of genes and other potentially relevant genomic elements, which is in part driven by the prospective uses of the sequence. Many of these stages require active interpretation and intervention in the production of data.

Thick sequencing has no fixed referent: it does not denote a particular event, process, object or project. It is, rather, a concept that encourages scholars of genomics to concentrate efforts on understanding those aspects of genomics that, as I will demonstrate, are as crucial to the products and processes of sequencing as the well-studied and crucial stages in which sequence reads are

generated and compiled in successive procedures to produce ever larger contiguous stretches of DNA sequence.[1]

In developing this distinction between thin and thick characterisations of sequencing, I am building upon Leonelli's (2016) work on data, which highlights the importance of understanding the processes involved in "packaging" data to enable it to be mobilised, integrated and employed by a variety of potential users. The distinction between thick and thin perspectives does not disrupt the centrality of data and data practices to our understanding of sequencing. Rather, the practices, collaborations, infrastructure and data that are included in the thick sequencing perspective are more varied, complex and networked than those associated with thin sequencing. Furthermore, reinterpreting sequencing as thick allows us to adopt the rich conceptual apparatus that has been developed to understand the epistemologies and pragmatics of data-centric or data-intensive science.[2]

In this paper, I provide a thick account of the mapping and sequencing of the genome of the domestic pig (*Sus scrofa*), encompassing more than just the production of 'raw' sequence. To quote Christopher Tuggle, a pig genome researcher at Iowa State University: "the sequence itself isn't very useful, we need to know where the landmarks are."[3] This concern with the usability of the sequence is allied with current policy directions in funding organisations that aim to improve and accelerate the translation of genomic data into (usually clinical) outcomes (e.g. Wellcome Trust, 2010; for the National Human Genome Research Institute, Green et al., 2011).

The pig genome sequencing work that I examine presents a well-resolved distinction between the thin and thick sequencing perspectives. For example, if we just take the determination of the order of bases, which was conducted between 2006 and 2009 at the Wellcome Trust Sanger Institute (in Hinxton, UK), then it looks centralised and automated. But considering the broader conception of sequencing provides a whole other picture in which a range of institutions contributed, over a longer timeframe, to both work preceding sequence determination and the development and processing of the Sanger Institute's raw sequence. This thick picture of sequencing includes the obtaining of DNA from several different breeds of pig, construction of four genomic DNA libraries containing clones of parts of this DNA, physical mapping, distributed revised assembly, sequencing genomic regions of particular interest in higher resolution, annotation and comparison with human and other species' genomes. This picture reveals a different organisation of the work and roles of particular skills.

Pig genome sequencing represents a genomics that took established organising principles and methodologies from prior projects, such as mice, cattle and – especially – human. The early stages of a new form of work involve a considerable amount of improvisation and trial-and-error. Early genome projects will therefore not necessarily be representative of sequencing once it became a

[1] In asking for further clarification on the thick-thin distinction, a reviewer asked whether 'what Incyte and Human Genome Sciences were doing with cDNA sequencing would be thin or thick sequencing,' referring to two private sector genomics companies established in the early-1990s. My answer would be that, as thin sequencing addresses a sub-set of operations encompassed by a thick sequencing perspective, these companies were conducting both thin and thick sequencing. Depending on one's scholarly interest in the workings of these companies, however, either a thin or a thick approach may be more pertinent.

[2] In particular, the perspective on data-centric science that has been developed by Leonelli (2016) and others (Stevens, 2013, for example, regarding genomics) that focuses on the active construction of the means to produce and circulate data, and the effect of such infrastructures on the status and value of particular data. This approach demands that understanding the role of data in data-centric areas of biology requires more than the circumstances of its immediate generation and eventual use.

[3] Christopher Tuggle, Skype interview with author, 3rd March 2017.

more established part of biological research (García-Sancho, 2012, especially pp. 21-64). Pig genome sequencing used the two dominant approaches to sequencing that arose out of the efforts to sequence the human genome: map-based (hierarchical) shotgun and whole genome shotgun (see figure 1). The map-based sequencing relied on a physical map produced from 2003 to 2005, and the whole genome shotgun data supplemented the map-based sequence data. Additionally, the pig genome community was and is relatively small and it is therefore possible to investigate most of those who were involved.[4] This helps me to avoid the reliance on accounts by prominent people based in large centres or ethnographic research conducted in those same centres, which methodologically structures a thin view of sequencing.[5]

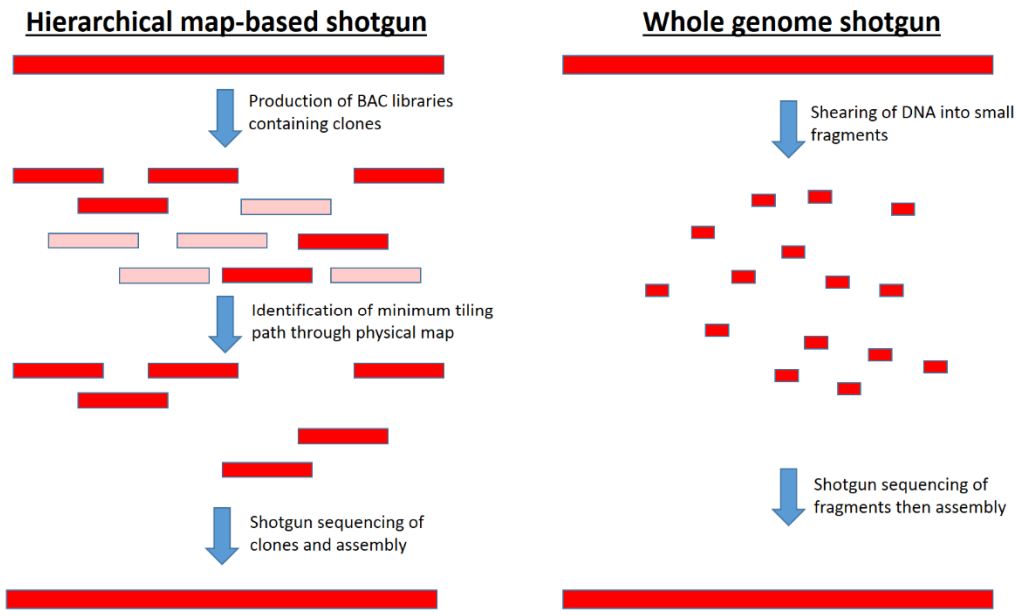

Figure 1 - A simplified depiction of the two chief approaches to genomics during and after the human genome project. On the left is the hierarchical map-based shotgun approach, which uses a physical map to produce a minimum tiling path to inform which Bacterial Artificial Chromosome (BAC) clones to sequence and consequently assemble into contigs. BACs and their yeast equivalent YACs are fragments of DNA sequences – clones – stored within the plasmids (circular DNA) of microorganisms. On the right is the whole genome shotgun approach in which the DNA is sheared into fragments, which are sequenced, and then assembled through high-powered computation to calculate the probabilities of overlaps between fragments.

In expounding upon the thick perspective, I outline an expanded view of sequencing. The explication of it in this paper concentrates, however, on only one aspect of genomics, the production of reference sequences. This is only one motivation for sequencing among many, and of one species among millions. The organisation of work and processes involved in the sequencing of other species, and for other purposes such as examining biological diversity, tracing evolutionary history, food testing, functional ecology and forensic investigation, may differ in significant respects from the

[4] It was certainly considerably smaller and more cohesive than the human genome community. Compared with the Swine Genome Sequencing Consortium, however, the sequencing consortia associated with the human and mouse genomes (for example) had different histories, organisation and composition, and therefore the data concerning their size and composition are not comparable. With colleagues who are working with data on human and yeast sequence submissions to the European Nucleotide Archive database, I am currently working on quantitative and qualitative analyses of data derived from pig sequence submissions to the same database, with a view to characterising the communities and networks involved in sequencing (but not necessarily whole genome sequencing) for each of the three species.

[5] See García-Sancho (2016), on a possible strategy to avoid this in historical research on human genomics.

account detailed in this paper. I make no claims for the representativeness either of the pig as the subject of sequencing, or of the production of reference sequences as its object. What I do aim to do is to demonstrate the power of the distinction I present and the possibilities opened up by taking a thick perspective on sequencing. Most pertinently, the thick perspective has the potential to stimulate an examination of all of the relevant practices and operations associated with sequencing conducted for different purposes, which would help to underpin more fine-grained comparative analyses between them.

This paper is based on archival research including on Alan Archibald's personal papers, documents and emails sent to me by key participants such as Lawrence Schook, examination of published materials and a series of oral history interviews that I have conducted with members of the pig genetics community and people who worked at the Sanger Institute.

I begin the paper with a discussion of the historiography of genome sequencing, before providing first an historical background to pig genomics, and then a detailed account of the sequencing of the pig genome. I demonstrate that the range of actors, practices and outcomes that the thick perspective covers is broader than those encompassed by thin sequencing. Throughout I will point to the historiographical and philosophical consequences of adopting a thick approach and including certain practices and processes in narratives of sequencing projects.

2. Historiographical background

The historiography of sequencing has been understandably dominated by the human genome project, and the practices, institutions, and actors associated with it. The human genome project lends itself to thin interpretations of sequencing due to the prominent role of the so-called G5 sequencing centres and the private company Celera.[6] As a result of its scale and salience, human genome sequencing has enabled the thin perspective to dominate scholarly interpretations of sequencing in general. The account of the sequencing of the pig genome in the rest of this paper is intended to help supplement and broaden the historiography that has been shaped considerably by human genome sequencing.

A common theme in accounts of the human genome project by scholars and participants alike was that this enterprise imported 'big science' and its associated characteristics into the biological sciences (Collins et al., 2003; Davis and Colleagues, 1990; Glasner, 2002; Hilgartner, 2013). Big science is characterised by the use of "large, expensive instruments, industrialization, centralization, multi-disciplinary collaboration, institutionalization, science-government relations, cooperation with industry and internationalization" (Vermeulen, 2016, pp. 199-200; see also Galison & Hevly, 1992). The effort to sequence the human genome has been compared to the Manhattan Project (Lenoir and Hays, 2000) and the US Space Program, in that "an immense, generalized capacity for technical action has been created" by the establishment and evolution of institutions, the training and deployment of personnel, and the development of techniques, instruments and protocols (Barnes & Dupré, 2008, p. 43). Sequencing the human genome, after all, involved large teams working towards an ambitious goal.

[6] I have chosen not to capitalise the words of the 'human genome project' to reflect that there was no such single organisational entity as the 'Human Genome Project' responsible for the sequencing, but a shifting collaboration of laboratories, centres, and funding and coordination initiatives.

Sequencing centres sought to industrialise the processes, and the focus was on the improvement of the efficiency of production and pipelines.[7] This implied a greater role for automation, standardisation and improving the flow from one part of the process to the next (Hilgartner, 2013; Stevens, 2011). There was not a uniform approach to sequencing the human genome, however, with two different approaches pursued by the 'official' public project and the main private sector initiative. The preference for hierarchical shotgun sequencing on the part of the 'official' human genome project allowed sequencing to be coordinated, with different centres sequencing different parts of the genome. This also permitted laboratories to try alternative methods and strategies (Bostanci, 2004, pp. 169-170). It therefore allowed different research interests and capabilities between laboratories and across nations differences to be accommodated.

As the human genome project proceeded, "automated machine rooms were established in a triumph of organization and routinization" (Barnes & Dupré, 2008, pp. 42-43). The automated sequencers were deemed crucial to requiring less human intervention and, later, to making sequencing more 'efficient' by being more cost-effective and requiring less (skilled) labour-intensive work (García-Sancho, 2012; especially pp. 131-143 and 163-168). An interpretation of this is that sequencing came to rely less on the labour of highly skilled scientists, and more on the routinised, standardised labour of technicians trained to operate machines developed and built by companies such as Applied Biosystems. Interestingly, however, rather than de-skilling sequencing, the establishment of large high-throughput genome centres has tended to foster the development of new skilled work. For instance, the sequencing process itself still requires highly skilled "careful laboratory work, testing, and judgment calls" (Stevens, 2013, p. 112). The work involved in creating and maintaining data infrastructures and pipelines also requires highly skilled teams (Leonelli, 2016).

The rapid automation of sequencing, and the remarkable development of newer machines able to sequence longer sections of DNA, often increasingly in parallel, increased the speed of sequence production, and contributed to a dramatic decline in the cost of sequencing.[8] This acceleration has led some scholars to characterise the human genome project as 'Fast Science' as well as 'Big Science' (Fortun, 1999). The speed and quantity of data production led many areas of the biosciences to develop means to store, manage, make accessible and make sense of the produced data. As a consequence, computers and other information technological infrastructure became central to the storage, transmission and management of the large amounts of data generated through sequencing work (García-Sancho, 2012; Stevens, 2013; Strasser, 2011).[9] In addition, software was developed to enable manual and automated approaches towards the handling, integration, analysis, comparison and interpretation of data.

In some areas, this has led to new forms of data-centric science, in which the practices, organisation of research and the formulation of knowledge claims are reshaped to make use of the large amounts of genomic information that has become available. Various scientific communities have developed

[7] With respect to the narrative of 'industrialisation,' Bartlett (2008), p. 99, comments that "The Human Genome Project appears, therefore, to be an island of Modernity in a perceived sea of post-Fordism, post-industrialism, and post-Modernism." However, as Stevens (2013), pp. 86-105, has shown, large-scale industrialised sequencing centres such as the Broad Institute are very much post-Fordist in their organisation of work and space.

[8] See, for example, https://www.genome.gov/sequencingcostsdata/

[9] In this paper I take no sides in the debate over the extent to which the importation of computers and information technological practices into biology has shaped the reconfiguration of biological research towards the production and handling of certain forms of data suited to those technologies and associated practices (Chow-White and García-Sancho, 2012; Lenoir, 1999; Leonelli, 2016; Stevens, 2013).

standards for the production, labelling and circulation of data, and its entry into data infrastructures such as databases and ontologies (Leonelli, 2016).

As a result of the nature of the organisation of the human genome project, sequencing has been conceived as an activity centred on large international collaborations. Despite this, the collaborations could also be characterised as more networked and decentralised (or more locally centralised on particular instruments) than projects in 'big physics', which typically require instruments that are orders of magnitude vaster in scale, cost and associated organisational complexity. Access to information infrastructures and the data contained therein is also widely dispersed and decentralised (Vermeulen, 2016, p. 204-205).

Changes in the attribution of credit have been associated with the development of large collaborations, most strikingly the dramatic increase in the number of authors listed on sequencing papers (Glasner, 2002). This is certainly the case for the pig sequencing that will be explored in this paper, with the thick sequencing perspective in particular revealing extra dimensions to international collaboration beyond the steering committee of the formal Swine Genome Sequencing Project through which the thin sequencing was coordinated. The thick perspective reveals more actors, and different regional organisational patterns among those actors. For instance, there were clusters of collaborative networks associated with the overall project that contributed particular aspects, for example the sequencing of cDNA to aid with annotation, a task performed by Japanese scientists.

Much of the historiography of genomics has been based on the sequencing of the human genome, due to the scale and political and cultural salience of this enterprise. However, examining other work beyond this will be important to furthering our understanding of the range, nature and development of genomics and sequencing. Examining the sequencing of other organisms can help us identify different ways in which the organisation and conduct of this work can occur (on yeast, for example, see Parolini, 2018). To that end, I first provide some background to genomic research involving pigs before detailing sequencing work involving those animals, and how its organisation differs from the large-scale, automated and centralised models of human genomics. As will be shown, a thick approach to assessing the sequencing of the pig genome allows one to elucidate different models of how genomics is and has been organised, an approach which may be applicable for genome projects involving other species.

3. Pig genomics – background

Collaborative projects to systematically investigate the pig genome began in the early 1990s. One project was based at the United States Department of Agriculture Meat Animal Research Center (USDA-MARC) and another was funded by the extramural Cooperative State Research, Education, and Extension Service of the USDA and took place largely in universities. In Europe, there was a collaboration funded by national agencies, ministries and the European Commission (the Pig Gene Mapping Project; PiGMaP). Initiatives also involved groups in Japan, Korea, Australia and Scandinavia. The aim was to identify genetic markers on each of the 20 distinct chromosomes of the pig and locate them, primarily to produce maps that could be used to advance the detection of Quantitative Trait Loci (QTL): areas of the genome linked to quantitative variation in phenotypic features such as fatness or meat quality. The idea was that once these were identified and mapped, breeders could use markers to select for or against traits with greater precision than previous livestock improvement efforts.

There are two main means of mapping QTL: genetic (or linkage) mapping and physical mapping. Genetic mapping allows one to ascertain the *relative* order of genes (or genomic markers) in linkage groups – i.e. areas of the chromosome whose genes tend to be inherited together (see figure 2). Physical mapping can ascertain the precise positions of genes and genomic markers on chromosomes (see figure 3).

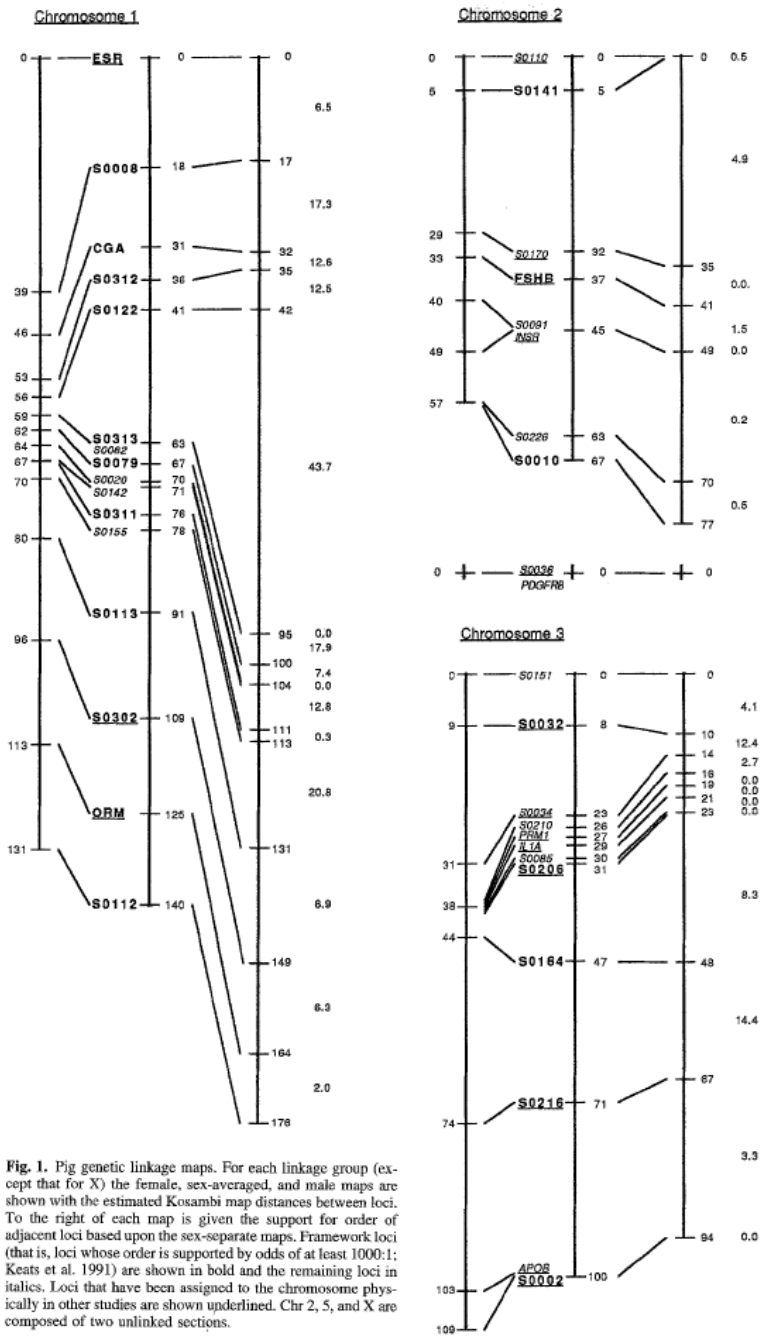

Fig. 1. Pig genetic linkage maps. For each linkage group (except that for X) the female, sex-averaged, and male maps are shown with the estimated Kosambi map distances between loci. To the right of each map is given the support for order of adjacent loci based upon the sex-separate maps. Framework loci (that is, loci whose order is supported by odds of at least 1000:1; Keats et al. 1991) are shown in bold and the remaining loci in italics. Loci that have been assigned to the chromosome physically in other studies are shown underlined. Chr 2, 5, and X are composed of two unlinked sections.

Figure 2 - Linkage maps for three pig chromosomes, taken from a 1995 paper authored by key participants in the European Commission funded PiGMaP consortium, together with collaborators outside Europe (Archibald et al., 1995).

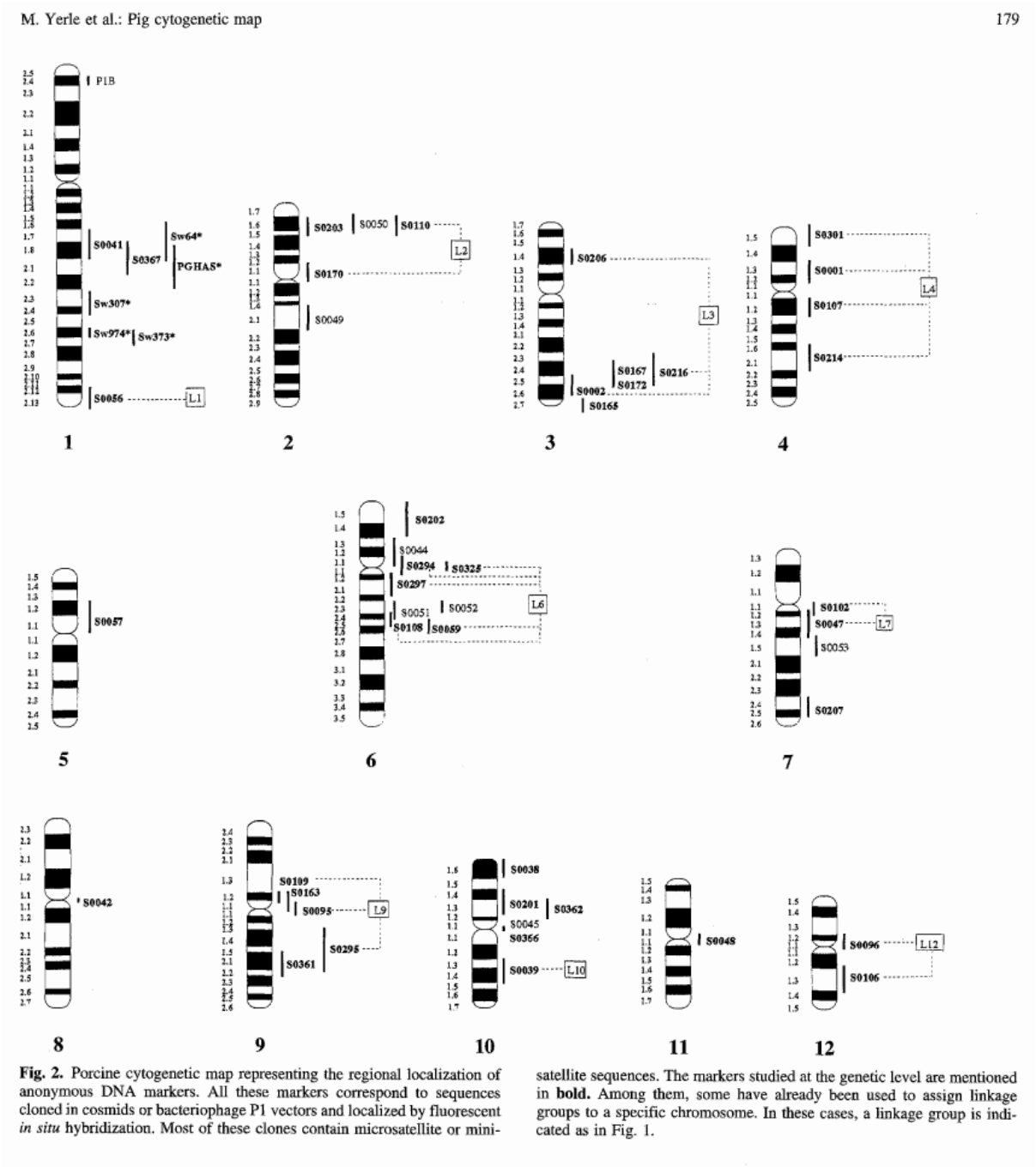

Fig. 2. Porcine cytogenetic map representing the regional localization of anonymous DNA markers. All these markers correspond to sequences cloned in cosmids or bacteriophage P1 vectors and localized by fluorescent *in situ* hybridization. Most of these clones contain microsatellite or mini-satellite sequences. The markers studied at the genetic level are mentioned in **bold.** Among them, some have already been used to assign linkage groups to a specific chromosome. In these cases, a linkage group is indicated as in Fig. 1.

Figure 3 - Physical map of 12 pig chromosomes depicting the exact positions of markers. In the early 1990s, 'physical' and 'cytogenetic' were both used for this form of work. The map is from a 1995 paper authored by key participants in the European Commission funded PiGMaP consortium, together with collaborators outside Europe (Yerle et al., 1995).

Genetic mapping primarily involved the identification of microsatellites (what are called type II markers), regions of repetitive sequences that are highly variable across individuals. Different breeds were crossed to maximise the chances of revealing polymorphisms, for example different lengths of repetitive sequence, to enable the degree of linkage between markers to be ascertained. The results were then integrated into databases and linkage relationships and groups identified using software developed and adapted for the purpose. The assignment of groups of linked type II markers to relative positions on chromosomes was aided by the physical mapping of predominantly type I loci to regions of the chromosome – type I loci are known genes linked to variation in particular phenotypic traits. Genetic and physical mapping therefore generated and related sets of markers

derived from different techniques. Where they produced markers in common, these would be used to integrate different maps (e.g. Rohrer et al., 1996).

The collaborative work required to make maps using genotyping data generated in different locations and to integrate the resultant maps helped to consolidate an international network of pig geneticists. From the 1990s, the members of this network met and coordinated their efforts at international events, such as at the annual Plant and Animal Genome conference and the meetings of the International Society of Animal Genetics, in addition to more regular contact between core members of the community by email, telephone and teleconference. The community was further nourished by the creation and distribution of genomic resources such as the primers made available by Max Rothschild, then pig coordinator of the USDA's National Animal Genome Research Program, and the IMpRH radiation hybrid panel produced by a collaboration between L'Institut National de la Recherche Agronomique (INRA) in France and Lawrence Schook's group at the University of Minnesota in the US.

In the early 2000s, this international community of pig geneticists sought to secure funds to produce a reference sequence of the pig genome. At first, this was under the auspices of an 'agricultural genome' first mooted in the early 2000s (e.g. Pool & Waddell, 2002). The agricultural orientation reflected the research agenda of much of the community and their prior and ongoing funding streams, which were concerned with producing data, tools and methods to be able to identify genes and other markers that could then be used in programmes of selective breeding in the livestock industry. Some of these researchers combined this work with an interest in the biology of domestication.

The community soon shifted its arguments, however, towards securing National Institutes of Health (NIH) funding. The NIH were looking to fund the sequencing of a mammalian genome to aid in the analysis of the human genome. A White Paper published by key figures in the pig genome research community therefore cited the genetic similarity between pigs and humans and emphasised the potential of the pig as a model for biomedical research (Rohrer et al., 2002). Although they were unsuccessful in acquiring NIH funding, the positioning of the pig as a biomedical model by members of the pig genome research community has continued (e.g. Groenen et al., 2012; Kuzmuk and Schook, 2011; Schook et al., 2005a).

The efforts to advance research into the pig genome built upon prior work on the human genome. Participants in human genome mapping and sequencing efforts also attended pig genome mapping meetings, for example Peter Marynen at a meeting of European pig genome mappers in Ghent in 1993 and Aravinda Chakravarti at the First International Workshop on Swine Chromosome 7 in Wisconsin in 1995 (Chakravarti, 1996).[10] Alan Archibald, based at the Roslin Institute near Edinburgh in Scotland and the co-ordinator of PiGMaP, served on the Co-ordinating Committee of the Medical Research Council-funded UK Human Genome Mapping Project (HGMP) in the 1990s.[11] The techniques and standards employed in human genome mapping were adopted and adapted by the pig genome research community, for instance by using the same kind of nomenclature and adhering to the Bermuda Principles, Fort Lauderdale agreement and the Toronto statement regarding the release of data (Archibald et al., 2010). The more comprehensive maps of human genes and

[10] Marynen is listed as a speaker in: Chris Haley (ed.) '4th EC PiGMaP MEETING 17-19 June 1993 HET PAND, University of Ghent, Belgium' report, in Alan Archibald's personal papers.

[11] 'MRC HGMP Co-ordinating Committee', page 27, G Nome News, Number 16, February 1994, edited by Nigel K. Spurr and originally published by the UK Human Genome Mapping Project. Cold Spring Harbor Laboratory Archives Repository, Identifier SB/9/2/54. Available online at: http://libgallery.cshl.edu/items/show/75888 Accessed 20th October 2017.

sequences of human DNA were used as an important basis of comparison for the pig mappers, and probes containing human DNA were used to identify markers in the pig genome. Comparisons took advantage of the evolutionary relatedness and conservation among mammals, and therefore the relative similarity of their DNA.

In September 2003, the community's efforts to coordinate the strategy and funding of pig genome sequencing led to the formation of the Swine Genome Sequencing Consortium (SGSC). The meeting to launch the SGSC was held at the INRA facility in Jouy-en-Josas near Paris and co-hosted by INRA and the University of Illinois Urbana-Champaign. The co-hosting duties re-capitulated the partnership that had produced the IMpRH panel, as Lawrence Schook had moved to the University of Illinois Urbana-Champaign from the University of Minnesota in 2000. After attempts to obtain funds from the NIH failed, funding for the SGSC was acquired from institutions that had previously sponsored pig genetic research. These were funders of agricultural rather than biomedical research and included the Biotechnology and Biological Sciences Research Council (BBSRC) and the Department for Environment, Food and Rural Affairs (DEFRA) in the UK; the European Union; and the US Department of Agriculture (USDA), the National Pork Board, the Iowa Pork Board, the North Carolina Pork Council, Iowa State University and North Carolina State University in the United States. On the award of $10 million dollars towards the sequencing of the swine genome, USDA Under Secretary for Research, Education and Economics, Joseph Jen, commented that "[b]y decoding the sequence of the pig genome, scientists can explore new ways to improve swine health and to increase the efficiency of swine production."[12] While biomedical applications were still hoped for, agricultural research priorities now dominated the sequencing project.

The SGSC had decided that the pig sequencing would be primarily map-based and hierarchical, with some additional whole-genome shotgun sequencing to provide some data for the final genome assembly. The first task was to construct a high resolution physical map of the pig genome. There was continuity in the personnel involved in previous projects to map the pig genome, and producing a high resolution physical map of the precise position of markers drew upon previous work in genetic, physical and comparative mapping.

A comprehensive physical map of the pig genome was produced, and clones from four genome libraries were sent to the Sanger Institute for hierarchical map-based shotgun sequencing. There was also therefore continuity between this new mapping work and sequencing. Sequencing understood thinly took place primarily at the Sanger Institute from 2006 to 2009. A thick approach to sequencing, however, allows me to expand the historical narrative beyond the Sanger Institute.

4. Pig genomic sequencing

A thick approach to sequencing provides the opportunity to identify all of the steps in a particular sequencing process. In doing so, there is also the potential to foreground the iterative and recursive qualities of sequencing. Therefore, while I will detail my account of the different aspects of thick sequencing in separate sections, the different processes can and in some cases do overlap. Furthermore, when one moves beyond specific projects such as the one that took place at the Sanger Institute, the linearity and unidirectionality of the sequencing process is challenged still further.

[12] Joseph Jen, quoted in 'USDA AWARDS $10 MILLION TO SEQUENCE THE SWINE GENOME', USDA News Release, Washington DC, January 13th 2006. Found in Alan Archibald's personal papers.

4.1. Elucidating a minimum tile path

An attention to the thickness of sequencing helps break down the distinction between sequencing and other forms of work such as gene mapping. Scholarly literature on genomics has paid close attention to the practices and conceptual inputs, developments and implications associated with both mapping (e.g. Gaudillière & Rheinberger, eds., 2004; Hogan, 2014; Rheinberger & Gaudillière, eds., 2004) and sequencing (e.g. Barnes & Dupré, 2008; García-Sancho, 2012), yet this work has still largely been partitioned. Here I include the production of a physical map of the pig genome in my thick discussion of the sequencing.

Funding for the production of a high-utility integrated physical map were provided from grants awarded to Alan Archibald at the Roslin Institute by the BBSRC, DEFRA, the private company Sygen and the Roslin Institute itself. Funds were available from 2003 to 2005 to enable the work to take place. Two programmes of the USDA also provided support. Archibald first approached the chief executive of the BBSRC with a proposal in August 2000, after consultation with figures in the USDA's Agricultural Research Service, prompted by the announcement by Wes Warren of Monsanto that the company were developing BAC contig maps for swine and cattle. Archibald and colleagues emphasised the importance of ensuring such maps were in the public domain, and detailed the potential uses of the data, including "improving the resolution of trait gene mapping" in part by being better able to characterise and then map Single Nucleotide Polymorphisms (SNPs) that may then be associated with variation in traits of interest.[13]

To aid with mapping efforts, researchers created first yeast (YAC) and later, bacterial artificial chromosome (BAC) libraries of cloned pig DNA in several laboratories around the world during the late 1990s and early 2000s.[14] Four BAC libraries were used in the construction of the high-resolution physical map used in the sequencing of the pig genome. Two (CHORI-242 and RPCI-44) were from the Children's Hospital Oakland Research Institute BACPAC Resources Center in the United States, led by Pieter de Jong.

The CHORI-242 BAC library was produced by Baoli Zhu from the DNA extracted from the white blood cells of a single Duroc (a North American domestic breed) sow named TJ Tabasco, who was born at the University of Illinois at Urbana-Champaign in 2001. The cloning was conducted according to a protocol developed in de Jong's laboratory (Osoegawa et al., 1998). The clones were inserted into a vector (a DNA construct) called pTARBAC1.3, and then *E. coli* cells were transformed to host the vector and the cloned pig DNA contained within. The CHORI-242 BAC library incorporates nearly 200,000 recombinant clones and was preferentially sequenced mainly due to its greater coverage of the genome (the overlapping DNA fragments contained in it were equivalent to 11 whole pig genomes). The other library developed in Oakland, RPCI-44, was funded by USDA-MARC and constructed by Chung-Li Shu. DNA for this was isolated from the white blood cells of four boars (each crosses of the Yorkshire, Landrace and Meishan breeds).

The third library, PigE BAC, was constructed and developed in the UK by Susan Anderson and Alan Archibald at ARK-Genomics, a unit of the Roslin Institute, and distributed from the Human Genome Mapping Project Resource Centre in Hinxton. The DNA was derived from the white blood cells of

[13] Alan Archibald, 17th August 2000, 'International Farm Animal Genome Projects', in Alan Archibald's personal papers.

[14] Initially, YAC libraries were developed due to the large numbers of recombinants required in BAC libraries. Due to problems with YACs, however, including their stability, chimerism and the presence of repeat sequences in the yeast genome, it was eventually decided to develop and use BAC libraries (Gary Rohrer, Skype interview with author, 30th March 2017).

male crosses between Chinese Meishan and European Large White pigs (Anderson et al., 2000). Finally, the INRA Porcine BAC library from Laboratoire de Radiobiologie et d'Etude du Génome (LREG) at INRA in France was constructed using DNA from the skin fibroblasts (connective tissue cells that synthesise collagen and other fibres) of a Large White male. The group was primarily interested in identifying retroviral elements, viral sequences incorporated into porcine DNA that it was thought could infect humans if pig tissues were transplanted into them for therapeutic purposes (Rogel-Gaillard et al., 1999). All four libraries involved the transformation of *E. coli* bacteria to host the libraries of clones.

The BAC libraries were sent to the Sanger Institute, that was contracted to perform the majority of the physical mapping work. Work was also conducted at The Keck Center for Comparative and Functional Genomics at University of Illinois at Urbana-Champaign under the auspices of the Livestock Genome Sequencing Initiative. Genoscope, the French national sequencing centre, sequenced the BAC-ends of the INRA BAC library.

The clones contained in the BAC libraries were digested with the restriction enzyme HindIII. The fragments thus generated were fingerprinted by electrophoresis on agarose gels.[15] This process involves running an electric current through the gel, separating the negatively charged DNA molecules according to size. Banding patterns produced and detected by a fluorimager as well as images entered into a fingerprint database were used as inputs into the software programme WebFPC to identify overlaps between fragments from different clones. Through using this programme, the 267,884 individual fingerprints were initially assembled into over 12,000 contigs, fragments containing unbroken stretches of base pairs. To reduce the number of contigs while increasing the average size, several procedures were used.

Firstly, sequences comprising an average of 707 bases at the end of each cloned fragment were determined (Groenen et al., 2012). These BAC-end sequences (BES) were deposited in the Ensembl and GenBank trace repositories, which stored raw data. Sequencing is therefore also a key part of this form of mapping, as it is with others. By aligning the BES with the human genome, using the database searching programme BLASTN, they were able to order them and thus merge contigs. This stage drew heavily on the established structural and sequence similarity between pigs and humans, and upon detailed prior studies of the synteny (the conservation of blocks of genomic order between two chromosomes) of pig and humans. Secondly, the statistical thresholds used in calculating the overlapping of clones and the merging of contigs could also be relaxed to merge the remaining contigs still further. As Alan Archibald put it to me, however, “you don’t want to produce a humanised pig genome,” so contigs were only joined if already supported by the fingerprint data.[16] Through these procedures and others involving the use of radiation hybrid maps, the initial thousands of contigs were reduced to 172, greatly increasing the contiguity of the map (Humphray et al., 2007).

Physical maps are important tools and resources in and of themselves. In pig research, they have been developed and used for the identification of QTL. Either the molecular basis of the QTL can then be investigated, or genetic markers situated close to the QTL identified. Pigs can then be genotyped for these and other markers, and these data can be used to inform which pigs to

[15] Not to be confused with the DNA fingerprinting developed by Alec Jeffreys of the University of Leicester, which is most famously used in forensic science and the determination of paternity. Jeffreys, incidentally, had some minor involvement in the early years of pig gene mapping. His colleague Esther Signer was a key participant and contributor to PiGMaP.

[16] Alan Archibald, interview with author, Roslin Institute, 17th November 2016.

incorporate in selective breeding programmes. Identification and mapping of QTL and associated genetic markers are key elements in practices aiming to effect phenotypic improvement in populations.

The physical mapping just described had its uses for these purposes, but was also an integral part of the overall project of sequencing. Through using "information about the extent of clone overlaps derived from the finger-print data and re-assessing the relative positions of paired BES alignments to the human genome," the physical mappers were "able to optimize the selection of an initial tilepath of minimally redundant clones through assembled clone contigs across the pig genome" (Humphray et al., 2007). They thus enabled the sequencers to identify the clones to be sequenced, optimised the sequencing operation (using the minimum number of clones) and helped to assemble the sequenced fragments.

Sequencing has been described as simply the production of an "ultimate map," more finely grained than is possible by genetic or physical mapping (McKusick & Ruddle, 1987; McKusick, 1991 and 1997). When physical maps are used as resources for sequencing, there is therefore no firm distinction between the work of sequencing and mapping; sequencing is a form of mapping. As historian Soraya de Chadarevian observes, in the same database for the nematode worm *Caenorhabditis elegans*, "a simple click on the mouse allows users to move from a locus on the genetic linkage map to its representation on the physical map and on to the sequence of the corresponding gene or, vice versa" (de Chadarevian, 2004, p. 95). This integration of different kinds of representation, as well as the value added to the earlier maps by the newer sequence data, however, came despite key differences between linkage and physical mapping and between mapping and sequencing that can be attributed to the different cultures, institutions and organisation of researchers involved. The different forms of mapping and sequencing are thus still considered by de Chadarevian to be distinct domains of activity with different products. A thick perspective enables one to encompass these different activities and cultures under one analytical umbrella.

4.2. Determination of 'raw' sequence

The thin sequencing perspective focuses on the determination of a 'raw' sequence of DNA bases. Frederick Sanger and his colleagues pioneered the sequencing of DNA in the 1970s.[17] 'Sanger sequencing' was the most prevalent technique used in sequencing before the development of 'next-generation' sequencing methods in recent years. It was based on the 'chain-termination' or 'dideoxy' technique. Initially, this was a laborious process that took a great deal of skill and time. From the early 1980s, efforts were underway to automate this process to improve the practicability of sequencing genomes of organisms.

[17] This is a simplified account of the history of DNA sequencing, for more historical and analytical detail, including the history of sequencing before the advances mentioned here, see García-Sancho (2012), pp. 21-64. Additionally, Onaga (2014) recovers the contribution of Ray Wu to early sequencing techniques.

Sequencing has historically depended on the sequencing of fragments of DNA rather than whole strands, and this is true currently, although methods to sequence whole strands have been proposed and are in development. Some automated sequencing machines can sequence longer fragments, though they are typically more expensive to use. This technical limitation means that approaches have had to be developed to sequence relatively small stretches of DNA at a time, and then integrate those sequenced stretches or fingerprints to produce ever larger and fewer contigs.

In hierarchical map-based shotgun sequencing, the chromosomes are cut up into pieces of around 100,000 to 150,000 base pairs, which are then inserted into BACs. The clones from these BACs are then cut up using enzymes, and the fragments are then sequenced. Finally, the order and location of the fragments is determined by an automated assembly method in which a computer programme identifies complementary sequences of DNA exposed at the end of the fragments produced by the enzymatic cutting. With the sequenced fragments placed in order, the sequence of the larger piece of chromosome is now known. In whole-genome shotgun sequencing, the genome as a whole is cut into small fragments, which are sequenced and then reassembled back into a whole genome. In the competing projects to sequence the human genome around the turn of the century, advocates of the map-based approach cited its accuracy, while partisans of whole-genome shotgun emphasised its speed (Bostanci, 2004, pp. 172-173; Brown, 2006, pp. 119-124; Wade, 2001, pp. 81-84). There were also deeper differences based on different organisational models and moral economies as well as different conceptions of the nature and structure of the genome which affected what partisans on either side saw as a feasible or valid approach (Bostanci, 2004, pp. 169-172).

The competition between two different models of sequencing demonstrates that choices, within particular material, social, disciplinary, political and policy constraints, have been made as to how thin sequencing is conducted. These choices have consequences in terms of the organisational and technical models by which they are realised.

In the case of the thin-focused sequencing for the pig genome conducted at the Sanger Institute, the basic approach outlined in a 2005 paper and described in more detail below was followed throughout (Schook et al., 2005b). The technical instantiation of that approach, however, changed over the course of the Sanger Institute's contribution to the project. The technology platforms changed, but also the organisation of the work, the latter inspired by the demands of sequencing the whole genome of an organism with 36 autosomes (non-sex chromosomes), with limited funding and time available. Chief amongst these changes to the organisation of the work was a shift to a stricter division of labour, and greater automation of certain tasks, including the use of robot colony-pickers.

Even at the point at which the pig genome was being sequenced, therefore, there was scope to automate previously non-automated tasks, and to institute changes to make the organisation of the sequencing work more like that of a factory than a traditional laboratory.[18] At the time of the pig genome project, the Sanger Institute research and development team would develop bespoke protocols appropriate to the genome being sequenced.[19] Since then, their aim has been to generate

[18] Stephen Hilgartner, writing on the human genome project, argues that the promoters of genome projects aimed to build large-scale specialised genome centres with factory-like organisation precisely to carve out a domain separate from molecular biology and genetics conducted in smaller-scale laboratories, so as to appear unthreatening to the existing organisational modes and moral economies in those disciplines (Hilgartner, 2017, especially chapter 4).

[19] This remains the case for organisms with genomes with potentially problematic properties, for example *Plasmodium falciparum.* This example and the information about the protocol for the pig was given to me by Carol Churcher, Head of Sequencing Operations at the Sanger Institute from 2008 to 2011, interview with author, Wellcome Trust Sanger Institute, 9th March 2017.

protocols and processes that are more generic and widely-applicable, and therefore standardised. Considered thinly, there has been a tendency in sequencing towards the greater standardisation of protocols and procedures, more factory-based organisation in which individual tasks are separated in space and conducted by individuals working only on that particular task and increased automation of particular tasks. However, these tendencies are uneven and partial. They reflect particular decisions, made on grounds of finance, policy, community standards and interests, disciplinary make-up, intellectual and practical aims, challenges and outputs, relationships and other factors. Manual work requiring experience and skill is interspersed throughout automated work. Sequencing, even considered thinly, is "an active process of extraction and construction shot through with difficult manual tasks and active judgment calls" (Stevens, 2013, p. 115).

Based on the experience of the physical mapping, the relationships that had developed, the fact that the Sanger Institute already had the clones and the ability of the USDA to fund work outside the US, the formal sequencing project was also to take place at the Sanger Institute. Several respondents described this as "logical," although according to correspondence and proposals dating from 2004 in Alan Archibald's personal papers, the possibility of having Baylor Human Genome Center in the US lead the project and conduct assembly and annotation of sequence data generated by them and five other centres was briefly considered.[20] To apprehend why the choice of the Sanger Institute as the sequencing centre with a different model to that of Baylor seemed logical, one therefore needs to turn back to the physical mapping, and so here the thick perspective is valuable. Once mapping is brought into the picture, the number of actors, institutions and practices multiply beyond the Sanger Institute and a centralised model based on it being the main sequencing centre for this project. The thin part of the sequencing was therefore dependent upon activities included in the thick interpretation of sequencing.

Following the model of the human genome project, the idea was to determine the sequence of base pairs using the BACs that were the shortest route through the physical map, the minimum tile path. This constituted 98.3% of the physical map. Once again, for this part of the process, corresponding to the 'thin sequencing', the four BAC libraries from the US, UK and France were used. In addition, a fosmid library (in which the DNA is inserted into a circular bacterial chromosome called an F-plasmid) was used, which incorporated DNA from the same sow used to construct the CHORI-242 library. Once again, clones from that library were preferentially used.

The (Sanger-method) sequencing was capillary-based rather than gel-based, which obviated the need for gel pouring or lane tracking (detecting the lanes in the gel was an extremely thorny problem for computers and therefore required human intervention to help the software read the gels). The DNA fragments pass through a capillary tube. The chain-terminating bases are tagged with a different colour depending on the base, and these fluoresce when hit with a laser. A camera records this and "traces" in the form of graphs of the four different colours are transmitted and recorded, and a computer programme detects the peaks and therefore assigns a base. Paired-end sequencing was employed, which meant that each fragment was sequenced from both ends.

In addition to the map-based approach, some whole-genome shotgun data were generated. Some of these data were incorporated into the assembly of the pig genome that was heralded with a paper published in *Nature* in 2012. The paper analysed the evolutionary implications of the data, and

[20] 11th January 2004, 'Proposed Hybrid Model for Swine Genome Sequencing', in 'Swine Genome Sequencing Project' folder, Alan Archibald's personal papers.

attempted to demonstrate the usefulness of the pig for biomedical research (Groenen et al., 2012).[21] The large number of authors (136, with 54 institutional affiliations) listed on the 2012 *Nature* paper seem to indicate that sequencing is indeed a more large-scale effort than previous modes of biological work. What is striking, however, is the extent to which the authorship of the paper reflects the involvement of many members of the pig genetics research community in many of the aspects of the initiation and coordination of the project, as well as involvement in the sequencing itself through the production of libraries, physical mapping, assembly and annotation; that is, if we are to understand sequencing from a thick perspective.

Following the determination of the order of base pairs and initial assembly, the genome then underwent further assembly and annotation. If one looked at sequencing thinly, the account would end here or early in the assembly section, rather than accounting for the fuller practices encompassed by assembly and annotation. Considering these produces a different picture of the topography and temporality of sequencing.

4.3. Assembly

The purpose of assembly is to build ever larger stretches of DNA from the sequence reads coming out of the sequencing machines. The average read length for sub-clones generated from each BAC was 707 base pairs. This clone-based sequencing generated 4X coverage, the equivalent of four whole genomes. The greater the coverage, the less likely that errors will make it into the final sequence assembly. A piece of software called Phrap was used to analyse the sequence data to assemble it into contigs. As well as the main body of the work at the Sanger Institute, some clones were also sequenced at the National Institute of Agrobiological Sciences in Japan, and some assembly work took place at The Genome Analysis Centre in Norwich, UK, after some Sanger Institute staff moved there following a strategic re-orientation at the Sanger Institute.[22]

With the stage of assembly into contigs complete, there were 279 contiguous pig clones. To further improve the quality of the genomic sequence, it had to undergo automated pre-finishing, gap closure and finishing by additional sequencing of selected BAC clones or genomic regions. The automated pre-finishing was accomplished by "primer walking" from the ends of contigs, by introducing a short strand of DNA called a primer with a sequence complementary to that to be determined at the end of the contig. From this a complementary DNA strand is synthesised, which is then itself sequenced. This enables contigs to be joined into fewer and larger DNA fragments, and

[21] A meeting of the Swine Genome Consortium in January 2011 reviewed 48 proposals for 'companion papers' to the main *Nature* paper. Of these, 12 were under the heading 'Application focused' and were oriented towards agricultural applications. 36 were under the heading 'Genomics Focused', most being concerned with further development of genomic data and resources and comparative and phylogenetic-style studies, with some biomedically-oriented papers included as well (Source: document dated 13th January 2011, 'Swine Genome Sequencing Consortium (SGSC) Genome and Companion Manuscripts Meeting' agenda, in 'Swine Genome Sequencing Project' folder, Alan Archibald's personal papers). In the end, 17 companion papers were published, of which 3 were wholly oriented towards agricultural applications, 2 directed towards biomedical applications, and the rest concerned with further development of genomic data and resources and comparative and phylogenetic-style studies, though with several of these being potentially relevant for agriculture and biomedicine (see: https://www.biomedcentral.com/collections/swine Accessed 10th July 2018).

[22] Over the period of Allan Bradley's directorship (2000-2010), moving "from sequencing genomes to using sequence data to answer important biological questions" (Wellcome Trust, 2005).

therefore to reduce the number of gaps in the sequence. After automated pre-finishing, 1,681 pig clones were contiguous.

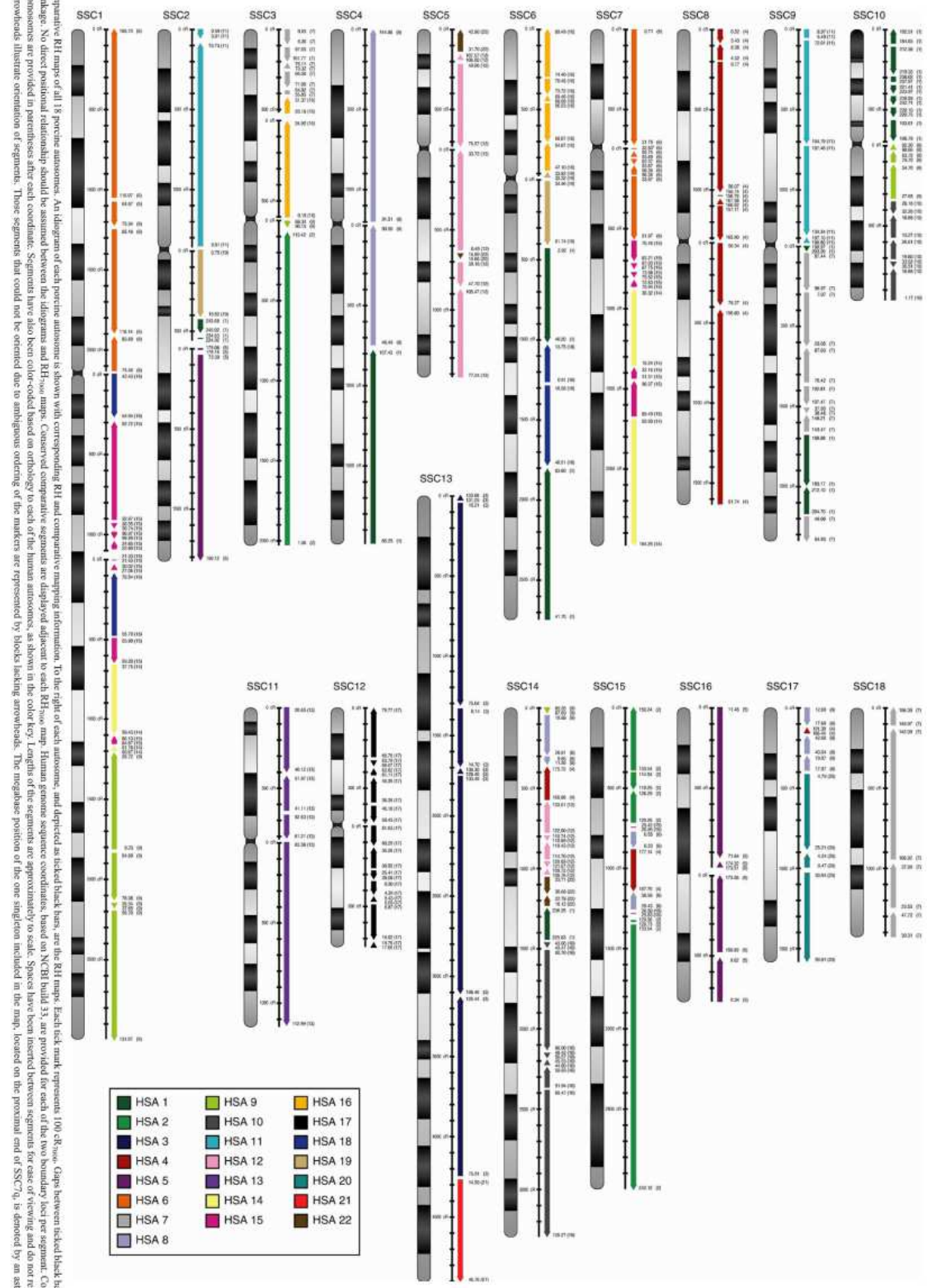

Fig. 1. Comparative RH maps of all 18 porcine autosomes. An ideogram of each porcine autosome is shown with corresponding RH and comparative mapping information. To the right of each autosome, and depicted as ticked black bars, are the RH maps. Each tick mark represents 100 cR$_{7000}$. Gaps between ticked black bars represent breaks in linkage. No direct positional relationship should be assumed between the ideograms and RH$_{7000}$ maps. Conserved comparative segments are displayed adjacent to each RH$_{7000}$ map. Human genome sequence coordinates, based on NCBI build 33, are provided for each of the two boundary loci per segment. Corresponding human chromosomes are provided in parentheses after each coordinate. Segments have also been color-coded based on orthology to each of the human autosomes, as shown in the color key. Lengths of the segments are approximately to scale. Spaces have been inserted between segments for ease of viewing and do not reflect gaps in linkage. Arrowheads illustrate orientation of segments. Those segments that could not be oriented due to ambiguous ordering of the markers are represented by blocks lacking arrowheads. The megabase position of the one singleton included in the map, located on the proximal end of SSC7q, is denoted by an asterisk.

Figure 4 - Comparative map depicting the 18 distinct porcine non-sex chromosomes, with equivalent parts of human non-sex chromosomes indicated adjacently (Meyers et al, 2005).

Assembly also requires checking that the sequence is substantially complete, and in addition to the automated methods it required judgement to do that. The judgement was informed by prior mapping efforts, including comparative maps detailing the correspondence between parts of the pig and human genomes (see figure 4). The Genome Evaluation Browser – gEVAL – produced and managed by the Genome Reference Informatics Team (GRIT) at the Sanger Institute was made available to the pig community to allow them to assess particular regions and suggest how to correct the assembly to improve it. Alan Archibald worked closely with GRIT to identify and correct errors. Evaluating and improving the quality of assembly is key to its potential use as data.[23] When a draft assembly was produced, Alan Archibald was able to check it using gEVAL. He scrolled through the assembly 2 megabases at a time and, comparing the orientation of the genes with a comparative map of the pig and human genomes. Archibald, when examining the screen (figure 5), would ask himself "does the pattern I'm seeing here fit with the expectations of that, if you like, that rough comparative map?" If something did not seem right, he would try to rearrange parts in different ways around, in his head or on scraps of paper, "but I'm not going to do that unless I've got some pig-specific information," ensuring that knowledge of the genetics of the pig disciplined the use of comparative data, "because I don't want this to be a human genome."[24] This was perceived to be a danger due to the extensive use of data and materials derived from human genomic research from the earliest days of the systematic mapping of the pig genome.

Figure 5 - Alan Archibald depicted in his office at the Roslin Institute. On the left screen he has a pdf document depicting the pig-human comparative map shown in figure 5. On the right screen he has the Sscrofa11.1 genome assembly open in a genome browser. He used the comparative map to identify equivalent regions in the human genome to the parts of the pig genome where gaps still exist, to indicate what may have caused problems in the assembly and thus identify which BAC clones to order and re-sequence. Photograph taken by author, 25th May 2017.

[23] Kerstin Howe, interview with author, Wellcome Trust Sanger Institute, 4th October 2017.
[24] The picture in figure 5 was taken from a video recording of Archibald taken by the author at the Roslin Institute on 25th May 2017. In the video, he is systematically working through gaps in a new sequence assembly. The purpose of the recording was to document usually undocumented scientific work, and to provide empirical materials concerning the role of comparison and homology in genomic research.

Other participants assisted in this work. For example, when at Roslin for a faculty sabbatical, Christopher Tuggle of Iowa State University noticed that there was something not right about the Sscrofa10 assembly while examining the interleukin-1 beta gene (*IL-1β*). Upon investigation, it was found that the programme used to assemble the BACs had not been written correctly. This meant that the way the algorithm was assembling didn't impart information about the orientation – the way round relative to the sequence being assembled – that the BAC should be in. The problem was thus identified (by Archibald) and the algorithm fixed.[25] There was therefore a need for expert manual judgement to assess the validity of the computational tools being used.

Organism	Name	Submitter	Date	Genome representation	Assembly level	Version status	RefSeq category
Sus scrofa (pig)	USMARCv1.0	USDA ARS	12/20/2017	full	Chromosome	latest	na
Sus scrofa (pig)	WTSI_X_Y_pig_V2	SC	02/08/2017	partial	Chromosome	latest	na
Sus scrofa (pig)	Sscrofa11.1 Synonyms: susScr11	The Swine Genome Sequencing Consortium (SGSC)	02/07/2017	full	Chromosome	latest	representative genome
Sus scrofa (pig)	Sscrofa11	The Swine Genome Sequencing Consortium (SGSC)	12/06/2016	full	Chromosome	replaced	na
Sus scrofa (pig)	WTSI_X_Y	SC	11/27/2016	partial	Chromosome	replaced	na
Sus scrofa (pig)	Tibetan_Pig_v2	Novogene	08/08/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Bamei_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Berkshire_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Hampshire_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Jinhua_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Landrace_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Large_White_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Meishan_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Pietrain_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	Rongchang_pig_v1	Novogene	08/05/2016	full	Scaffold	latest	na
Sus scrofa (pig)	ss10.2_mar2013	F. Hoffmann - La Roche AG	09/16/2015	full	Scaffold	latest	na
Sus scrofa (pig)	minipig_v1.0	BGI-shenzhen	03/18/2015	full	Scaffold	latest	na
Sus scrofa (pig)	Tibetan_Pig_v1.0	Novogene	02/06/2015	full	Scaffold	replaced	na
Sus scrofa (pig)	SscrofaMinipig	GlaxoSmithKline	01/10/2013	full	Contig	latest	na
Sus scrofa (pig)	minipig_v1.0	BGI-shenzhen	11/30/2012	full	Contig	replaced	na
Sus scrofa (pig)	Sscrofa10.2 Synonyms: susScr3	The Swine Genome Sequencing Consortium (SGSC)	09/07/2011	full	Chromosome	replaced	na
Sus scrofa (pig)	Sscrofa10	The Swine Genome Sequencing Consortium (SGSC)	05/19/2011	full	Chromosome	replaced	na
Sus scrofa (pig)	Sscrofa9.2	The Swine Genome Sequencing Consortium (SGSC)	02/23/2010	full	Chromosome	replaced	na
Sus scrofa (pig)	Sscrofa9	The Swine Genome Sequencing Consortium (SGSC)	11/02/2009	full	Chromosome	replaced	na
Sus scrofa (pig)	Sscrofa5	The Swine Genome Sequencing Consortium (SGSC)	07/11/2008	partial	Chromosome	replaced	na

Figure 6 - Genome assemblies for the pig submitted to GenBank. As GenBank is based in the USA, the date format is MM/DD/YYYY. Table adapted from the GenBank website: https://www.ncbi.nlm.nih.gov/assembly/organism/9823/all/ Accessed 09/07/2018.

As of July 2018, there are 25 genome assemblies that have been submitted to the publicly-accessible database GenBank. They are categorised in a number of ways, and have been submitted by multiple groups. Seven have been submitted by the SGSC. Other submitters include BGI-Shenzhen (formerly the Beijing Genomics Institute), the Sanger Institute ('SC' in the table), a genome sequencing company called Novogene that has conducted sequencing in China with collaborators from universities and has links with the Chinese Ministry of Agriculture, and two pharmaceutical companies, Hoffman-LaRoche and GlaxoSmithKline. The most recent submission comes from the USDA. Most of the submitted assemblies are full representations of the genome, meaning that the

[25] Christopher Tuggle, Skype interview with author, 3rd March 2017.

data were acquired from the whole genome, rather than just a part of it, but with different levels of assembly and assigned status. The assemblies submitted by the SGSC and the USDA are chromosome-level assemblies, meaning that there is a sequence for at least one chromosome. This sequence may still contain gaps. The most recent SGSC submission, which the National Center for Biotechnology Information that runs GenBank has selected as the representative genome for the pig, is not however designated as a complete genome. That designation would require that all chromosomes be sequenced without gaps in the sequence, and fulfil other criteria that will be discussed below. The other two assembly levels listed in the table are scaffold and contig. A contig is a continuous sequence in which there is a high confidence level in the order of the bases. A scaffold is a section of sequence that incorporates more than one contig, together with the gaps of unknown sequence known to exist between them. The aim of sequencers is to reduce the number of gaps, and therefore the number of contigs and scaffolds, and also to localise and place the scaffolds on the chromosome. To qualify for complete genome assembly level, the sequence must have no unlocalised or unplaced scaffolds. The following table shows how the submitted assemblies have changed over time for the SGSC submissions.

Name (date submitted)	***Coverage***	***Number of chromosomes***	***Total sequence length***	***Gaps between scaffolds (number of scaffolds)***	***No. of contigs***
Sscrofa5 (11.07.2008)		*10*	*813,033,904*	*1,584 (1,585)*	*44,057*
Sscrofa9 (02.11.2009)	*4X*	*19*	*2,262,579,801*	*3,133 (3,133)*	*101,117*
Sscrofa9.2 (23.02.2010)	*4X*	*19*	*2,262,484,801*	*3,116 (3,135)*	*101,112*
Sscrofa10 (19.05.2011)	*24X*	*21*	*2,772,757,746*	*3,915 (8,519)*	*266,137*
Sscrofa10.2 (07.09.2011)	*24X*	*21*	*2,808,525,991*	*5,323 (9,906)*	*243,033*
Sscrofa11 (06.12.2016)	*65X*	*19*	*2,456,768,445*	*24 (626)*	*705*
Sscrofa11.1 (07.02.2017)	*65X*	*20*	*2,501,895,775*	*93 (705)*	*1,117*

Figure 7 - Table providing figures for certain key measurements of successive genome assemblies submitted to GenBank by the Swine Genome Sequencing Consortium. The row for Sscrofa5 has been highlighted in grey as it is only a partial assembly. In this table I have used the DD.MM.YYYY date format used in Europe and much of the rest of the world.

The statistics can be confusing, because the assemblies are not necessarily directly comparable: for example, the number of chromosomes sequenced and assembled may not be the same. Although they are not equivalent in length, Sscrofa11 has a considerably smaller number of scaffolds and gaps between scaffolds compared with Sscrofa9, and the same is true for contigs. This is an improvement, as the fewer numbers of scaffolds or contigs there are, the greater confidence we can have that the

assembly is correct. More contigs and more scaffolds mean a greater likelihood of mistaken placement. Despite greater coverage of the genome (the number of reads of any given nucleotide in a sequence), Sscrofa10 has a higher number of scaffolds and gaps between scaffolds. This does not mean that the assembly is of a lower quality, but that extra chromosomes and extra-nuclear DNA have been included in the assembly. In particular, it includes an assembly for the Y chromosome which, to give an example in the Sscrofa11.1 assembly, contains 69 of the 93 gaps between scaffolds across all chromosomes. The Y chromosome notoriously contains many repetitive sequences that are consequently difficult to assemble. Any assembly including the Y chromosome is therefore likely to have its metrics negatively affected.

Although a representative genome is designated based on coverage and assembly statistics (including those relating to gaps and error rates) there is not any one complete or final sequence.[26] There are corrections to existing assemblies. There are numerous sequences of different breeds and, although the SGSC assemblies show greater quality over time, the standard of what constitutes a gold-standard assembly also changes over time. If we examine sequencing from a thick perspective, we can therefore qualify prior historical accounts that take the completion of a determined raw sequence as the end-point of genome projects in two ways. Firstly, by including the work done on a particular sequence assembly beyond the initial stages of assembly, for example through later stages and iterations (for the 'same' genome) of assembly and improvement. Secondly, by investigating sequencing activities separate from the generation and development of reference genomes.

Sequencing understood thickly is not only concerned with an improvement in the statistics over time for a representative genome. This is shown by the Novogene submissions that are sequence assemblies for different breeds of pig to the SGSC assemblies based primarily on a Duroc sow. There is an interest in the sequences of different breeds, and, for the purposes of animal breeding genetics, an acute interest in the variation in sequences. Therefore, we may anticipate that new needs will arise for which new categorisations and statistics will be generated for assemblies to be judged against.

Sequencing and assembly is an open-ended process, which involves the periodic submission of sequences and assemblies to databases like GenBank. At every stage, decisions are made of what to sequence (for instance, the breed), what part of the genomes receive particular (or little, or no) attention, the coverage, the sources (particular individuals, particular BAC libraries), the method, the machines used (which can vary in technique and chemistry, and operation), and how the assembly is conducted.

For the assembly, different software can be used, and decisions are made about the statistical confidence levels. Lower the stringency of these, and one can reduce the number of contigs, but at the price of an increased likelihood of assembly errors. Additional coverage and better techniques can aid the assemblers in reducing the likelihood of errors. Also, access to a high-quality physical map for the species they are working with, in conjunction with sequence data from related species with known synteny can be vital aids in assembly, even if they are time, labour and cognitively demanding. A consequence of the foregoing argument is that there is no *a priori* point at which we can say that sequencing ends. This point was echoed by Kerstin Howe of GRIT, who reflected that "there is always something to correct with a genome assembly, it's never done; it's only abandoned at a certain point," for instance because particular quality targets have been reached or resources

[26] Similar points have been made by Adam Bostanci with regard to the human genome project, on the publication of two versions of genomes produced using different methods and organisational models (2004), and the problematic of acknowledging and accommodating intra-specific sequence variation (2006).

have run out.[27] We may wish to define sequencing as the activities that occur under the aegis of a particular project, but this would be problematic unless we were to explicitly acknowledge that a study based on this definition is one of a particular sequencing project, rather than sequencing more generally. To answer questions like what the Sanger Institute sequenced or where the pig genome was sequenced requires shifting from a thin to a thick perspective, as only then can the constellation of inputs (such as BAC libraries and physical maps), outputs and decisions (concerning strategy and the division of labour) be captured.

In addition to published genome assemblies, there are sequence data submitted to DNA Data Bank of Japan, GenBank and the European Nucleotide Archive (ENA).[28] These are primarily sequences of chromosomal regions relevant to particular research. Rather than being superseded by a published, complete assembly, some of these sequences may still serve as reference sequences for specific areas of research. This can be because of previous sequencing of a defined region being of greater quality than the overall builds, or because of resequencing after the initial builds. An example of the former is the sequence of the swine Major Histocompatibility Complex, which occupies a region on chromosome 7 (Renard et al., 2006). For these researchers interested in the porcine immune system, the reference sequences were not those generated at the Sanger Institute under the auspices of the SGSC. Examples of *de novo* sequencing include the next generation sequencing of the DNA of eight breeds of domesticated pigs and wild boars from across Europe and Asia, the data from which was used in an investigation of the evolution and demography of the pig (Groenen et al., 2012).[29]

4.4. Annotation

In this paper, I have provided an account of the thick sequencing perspective through detailing the production of a sequence capable of wider travel and use. Annotation is a key part of making the assembled sequence capable of this. It is the process of attaching contextual information, for instance by identifying and assigning genes, to particular parts of a sequence assembly.[30] Without annotation, the sequence data is "by itself neither informative nor particularly interesting;" information needs to be attached to it in order that it may be able to circulate and be incorporated into the research of a variety of potential end-users (Nadim, 2016, p. 505; see also Leonelli, 2016).

The way in which the annotation takes place is not uniform across different genome projects. The assignment of the task may differ, as may the precise balance of automated annotation and manual annotation. If there is the time and resources to do it, manual annotation is preferred, but sometimes where these are lacking, automated annotation may be the only option.

---

[27] Kerstin Howe, interview with author, Wellcome Trust Sanger Institute, 4th October 2017.

[28] These three databases synchronise new and updated data submitted to each, and together comprise the International Nucleotide Sequence Database Collaboration. There are hundreds of thousands of individual submissions of sequence data of varying lengths to publicly-accessible sequence databases. Furthermore, sequence data that are not publicly-accessible will also likely be held, for example in private repositories for proprietorial reasons.

[29] The eight breeds of pig were Duroc, Hampshire, Jiangquhai, Landrace, Large White, Meishan, Pietrain, and Xiang. The wild boars were from Sumatra, Japan, two locations in the Netherlands, France, Switzerland, South China and North China. The study accession number at the European Nucleotide Archive is PRJEB1683. Other sequence deposits for pigs and closely related species are listed in Groenen (2016).

[30] This is what structural annotation consists of. Functional annotation involves attaching meta-data to structural annotations, and therefore depends upon this initial form of annotation.

As well as augmenting automated annotation, manual annotation can feedback into the evaluation and improvement of an assembly. For instance, if a known gene is not located in the process of manual annotation, this suggests a possible mis-assembly and therefore highlights a potential region that can be re-evaluated and corrected.[31]

To refine automated annotation approaches the ongoing contribution of data is required, for example the tagging of certain transcripts (Harrow et al., 2014). Automated annotation therefore requires informed manual work in order to function, and to be maintained and improved. It also requires people in relevant communities to identify the need for particular software, to guide and validate the development of that software, and to determine and discipline the particular inputs and outputs of the software operation. Software that is being developed, including the ongoing development of the Ensembl genome browser, involves constant interaction between the developers and the research community. The browser was developed by the Ensembl project, which was initially a joint initiative of the Sanger Institute and the European Molecular Biology Laboratory's European Bioinformatics Institute (EBI, which hosts the ENA, based on the same site as the Sanger Institute). The Ensembl project team now works wholly within the EBI. The project was created to formulate methods and tools for automated annotation, and the browser shows visualisations of sections of the genome with details of genes and other potentially relevant information shown parallel to the parts of the genome with which it has been associated through annotation.

In the pig genome project, several computational tools were used for annotation, including: scans for sequence patterns; mapping of pig protein sequences (acquired from public databases) to the genome; processing and alignment of cDNAs and Expressed Sequence Tags (ESTs) – many of which for the pig were generated by groups in Japan based at the National Institute of Agrobiological Sciences, Animal Genome Research Program and Japan Institute of Association for Techno-innovation in Agriculture, Forestry and Fisheries – downloaded from GenBank to the genome, and alignment of RNA sequencing data (RNA-Seq) to the genome. Redundancy was then removed, the set of genes was screened for pseudogenes (similar but non-functional versions of genes) and Stable Identifiers (identifying codes) were assigned to the genes and other elements of interest (Groenen et al., 2012).[32]

In addition to the automated annotation, the manual annotation team then at the Sanger Institute (and now based in the EBI), HAVANA (Human and Vertebrate Analysis and Annotation), provided additional support to the pig project. Jim Reecy of Iowa State University interested the HAVANA team in manual annotation of the pig genome when he spent a sabbatical there. As HAVANA did the work at no extra cost to the SGSC, their approach was to supply the pig genetics community with the tools and training to be able to annotate the genome themselves. In July 2008, the team at the Sanger Institute organised a workshop (a 'jamboree') to train scientists associated with the SGSC on how to annotate.[33] Targeted annotation aimed at regions of particular interest to researchers was pursued. In the case of the assemblies that took place during the initial swine genome sequencing

[31] Jane Loveland, interview with author, Wellcome Trust Sanger Institute, 4th October 2017.

[32] This is a simplified and abbreviated account of the annotation process. See the Supplementary Information of Groenen, et al. (2012) for a fuller account.

[33] The annotation 'jamboree' was pioneered by Celera Genomics and the Berkeley Drosophila Genome Project to annotate the *Drosophila melanogaster* genome. The July 2008 jamboree mentioned here was more of a lengthy training course than the "frontal charge on the genome" represented by the kinds of jamborees inaugurated by the fruit fly collaboration. The pig genome annotation was more like a combination of the factory and cottage industry models of annotation discerned by bioinformatician Lincoln Stein (2001), pp. 500-501, comprising automated pipelines and distributed smaller-scale annotation by researchers in their own laboratories, respectively.

project, annotation was conducted "on the fly," as Craig Beattie (a member of the SGSC then based at USDA-MARC) put it to me, while the assembly was ongoing, as well as after it was complete.[34]

There was additional training organised by one of the working groups that focused on annotating genes relevant to the immune system. Many of the genes were annotated by leading members of the Consortium in collaboration with the HAVANA group. The immune system genes were distributed to multiple individuals, who together comprised the 'Immune Response Annotation Group.' Researchers based in 13 institutions across 6 countries (China, France, India, Italy, Japan, UK and USA) had manual annotations attributed to them in this project. They used the manual annotation software Otterlace (developed at the Sanger) to annotate 1,300 genes, including confirming automated annotations (Dawson et al., 2013).

The annotated sequences that are produced through the automated and manual processes that form the Ensembl pipeline are published in the Ensembl database. Additional manual annotation is published on the HAVANA-led Vertebrate Genome Annotation (VEGA) database, which is built on the Ensembl database. Using the browsers, researchers can therefore access the annotated genomes (see figure 8), and additionally make comparisons between regions of genomes (for more on VEGA, see Harrow et al., 2014).

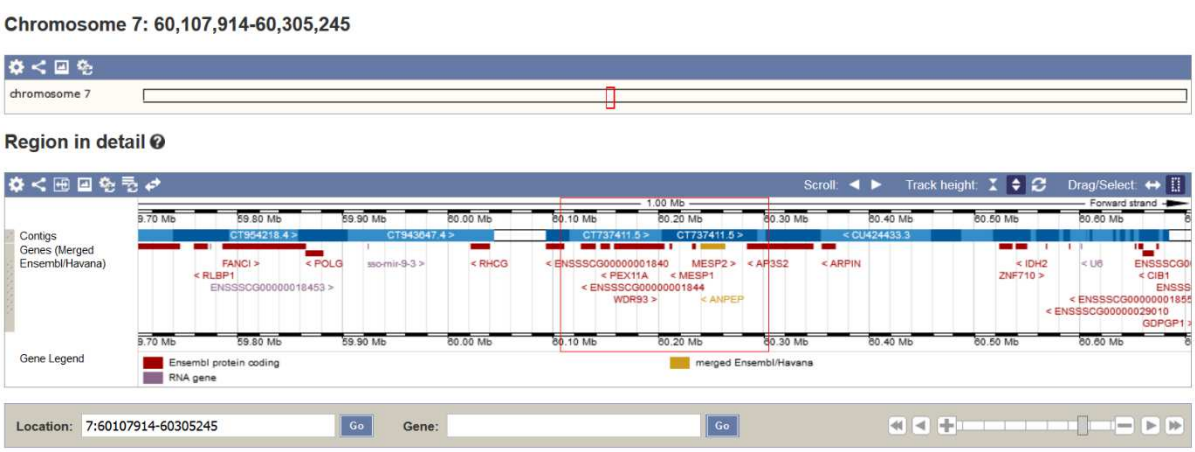

Figure 7 - Annotated region of chromosome 7 of Sscrofa10.2 in the Ensembl browser. This is just one part of the visualisations depicted on the browser page for this region. This summary displays genes and contigs. The more detailed one is capable of depicting multiple tracks pertaining to different kinds of data, and allows the user to zoom in until the order of bases in the sequence can be shown. This particular image is obtained from: http://may2017.archive.ensembl.org/Sus_scrofa/Location/View?r=7%3A60107914-60305245 Accessed 20.10.2017.

The way annotation works in practice shows that, as with assembly, automated processes form only one part of this work that is included in the thick sequencing perspective I'm proposing. The choices to use, adapt or develop particular automated procedures are made by skilled practitioners, usually in ways appropriate to the kinds of data available for a particular organism. Until recently, specific operating procedures have been used for each organism sequenced at the Sanger Institute. Although more standardised protocols are now being developed and used, judgements still need to be made about what annotation tools and strategies should be used to complete the sequencing of particular organisms.

In the case of annotation and sequencing in general looked at from a thick – rather than thin – perspective, the following historiographical consequences are posed: 1. sequencing still requires considerable tacit knowledge, and expert (discipline-specific) interpretation and input; 2. it is not

[34] Craig Beattie, Skype interview with author, 23rd March 2017.

necessarily centralised, top-down organised or fast/accelerating; and 3. automated processes are only one part of the activity of sequencing. Even the 'black-boxes' of sequencing machines or software may be better characterised as 'grey-boxes', because they are not fully closed; there is constant dialogue between the relevant biological research communities and manufacturers, and a choice of machines with different capacities and capabilities (and prices).

The thin perspective captures one stage of sequencing. As the preceding account of pig sequencing demonstrates, however, the work emblematic of a thin focus on sequencing – the determination of the raw sequence – relied on practices that both preceded and succeeded it, and that only a thick perspective shows. These include the development of libraries of clones, physical mapping using those clones to generate a minimum tile path, sequencing of selected clones and then iterative stages of assembly to close gaps. Finally, the sequence required annotation in order to be of use – and tailored to – different biological communities. Throughout this process, constant comparisons were made with data from the human genome, and other resources such as cDNA, ESTs and RNA-Seq data generated by groups across the world were drawn upon.

The pig project drew upon the methods and organisation pioneered in the human genome project: in particular, the map and clone-based approach. Pig genomics enabled the Sanger Institute to accentuate the ongoing drive to further improve the efficiency of the sequencing process through increased automation and greater division of labour. There was continuity in staffing. Jane Rogers, who had been Head of Sequencing at the Sanger Institute since 1993, was involved in the planning and early stages of the pig sequencing. She was replaced by Carol Churcher, who had also been at the Sanger Institute since its founding. Many of the key figures in the pig genetics community who participated in various stages of sequencing had been involved in prior mapping projects, such as Alan Archibald, Denis Milan and Lawrence Schook. There was even continuity in the source of libraries – Pieter de Jong's group was a source for the human project as well as the pig one. We may thus consider this analysis of pig sequencing to also be relevant to analyses of human genomic sequencing that preceded it, both the public and the private arms.[35]

[35] Although the private Venter-led initiative famously featured a whole-genome shotgun approach, it also used maps and data from the clone-based sequencing that the publicly (and charitably)-funded effort pursued.

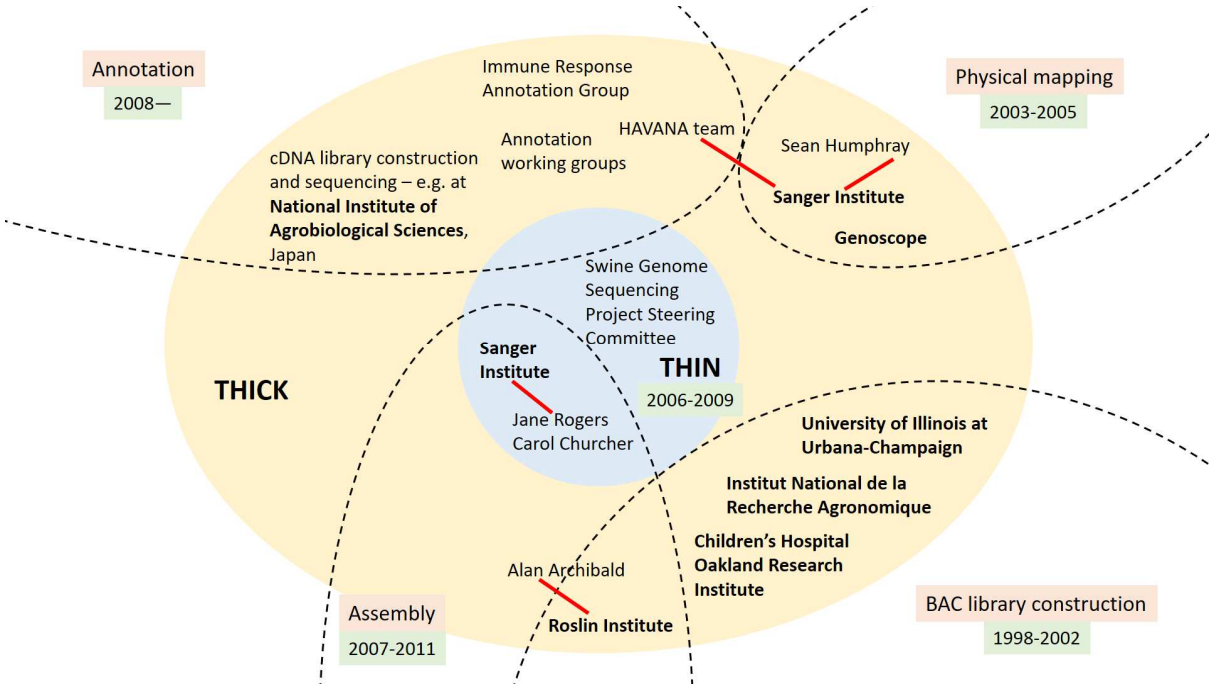

Figure 9 - Diagram illustrating the principles of the distinction between thin and thick sequencing, through an illustrative but not exhaustive or necessarily representative depiction of the institutions (in bold) and individuals or groups involved in different activities related to the sequencing of the pig genome. Red lines indicate institutional affiliation, the dates given are those for the activities associated with the production of the Sscrofa10.2 assembly, and black dotted lines indicate involvement in activities.

5. Conclusions

Through detailing an account of pig genome sequencing using a thick sequencing perspective, this paper has demonstrated that sequencing can be understood as a process that is open-ended: spatially, temporally and intellectually. Sequencing as an ongoing process involves the creation of libraries and maps, the working and workings of automated sequencing machines, and associated decision-making related to the use of them. The process also involves the assembly of the sequence, the development and improvement of statistical and computational tools, of chemistry and machinery, annotation, extra sequencing of certain parts of the genome, improvement of the contiguity and quality of the data, new reads, uploading, circulation and interpretation of data, management, curation and maintenance of data and data infrastructures.

There may be instances at which start or end-points of the sequencing process can be ascertained. One might conceive the approval by a body such as the Wellcome Trust, USDA or NIH of a proposal to sequence a particular species as a start. The initial receipt of clone libraries at sequencing centres, and the first entry into automated sequencing machines, may both be conceived as starting points. Or one might wish to identify how the tools, organisational capacity and desire by the community to sequence some or all of the genome came to be. The endpoints might be a published paper, or the online publication of a completed sequence.

Yet these starting points and endpoints are less discrete and definitive than on first inspection. In culinary terms, genome sequencing is more like cooking a perpetual stew in which ingredients can be added and the pot kept constantly on the boil, never fully complete. Firstly, the product is almost always incomplete. Gaps may remain, and there remain some errors in final published sequences. Secondly, either the product is an abstraction (purporting to be a reference sequence for a species, breed or strain where there is known to be genomic variation) or the product incorporates (or is

built to allow the incorporation of) genomic variants such as single nucleotide polymorphisms (SNPs). The former is not definitive and is therefore subject to contestation and revision, the latter can never be definitive.[36] As well as not being able to (historiographically or philosophically) privilege one stage of the sequencing process over any other, it is not possible to determine *a priori* the start and end points of sequencing.

Any attempt to methodologically or epistemically delimit sequencing therefore requires a specific historiographical (or philosophical) basis, and the limitations of this choice need to be acknowledged and used to inform any conclusions drawn. In the case of pig genome sequencing, an attempt to reduce sequencing to thin sequencing precludes one from understanding or appreciating many of the key decisions and research directions, especially concerning the purported 'logic' of the location and strategy of the thin-centred sequencing. The Sanger Institute was chosen – and seemed 'logical' – because significant parts of the physical mapping work had been conducted there, the clones were already there, the conducting of human genome sequencing there and the adoption of the human model by the pig community, and the relationships that had been established. So even to understand the objects of a thin perspective of sequencing, one must invoke the work and actors revealed by the thick perspective. In its attention to the production of a sequence with added value and usability, the thick perspective will also allow us to apprehend how genomic research may contribute to strategic policy directions concerning translation. It also helps one to recognise key differences between institutions and their consequences, for example of the production of sequence at institutions that devote different levels of resources to adding value to the sequence through comprehensive evaluation and annotation.

An attention to the thickness of sequencing leads one to characterise the geography, the temporality and the nature of sequencing work in a fundamentally different way than for the thin perspective. Understood more thickly, sequencing takes longer, has less well-defined start and end points, is more institutionally-diverse, involves a plethora of different skill sets and background knowledge, and involves considerably more actors in general. A thick examination of sequencing reveals the active interpretation, intervention, assessment, evaluation and creativity of scientists. It requires an appreciation of the relationships between scientists, technical staff, project managers, administrators, industries and funders. Throughout the sequencing process detailed in this paper, there was an interplay and interpenetration between adapting and refining protocols and processes and using standardised tools and procedures. Where elements of work have been automated, the manual, creative and interpretive work of scientists may still be required both in and around the automated processes. These scientists work in the processes to evaluate, maintain and refine them, and around them to take advantage of the 'black-boxing' in order to concentrate on new problems. In sequencing interpreted in a thicker manner, some of the features of this reconfiguration may be discerned in the apparently centralised work conducted in massive genome centres. For example, in the pig genome project described above, the development of the principles and processes of assembly and annotation had culminated in the use of automated pipelines, yet there was still room for manual intervention both in the later stages of assembly and also in annotation.

For automated sequencing, lower costs have made the geography and concentration of it more diffuse and less centralised. Citing the sequencing services offered by shared facilities in research institutions as enabling sequencing to be "reconfigured as a small-scale, slower and artisanal form of work, subordinated to concrete research necessities," García-Sancho observes that "other

[36] See Bostanci (2006) for a discussion of the notion of 'the human genome' and its supersession by the investigation of 'human genomic variation'.

sequencing is possible" (2012, pp. 176-177). Even thin sequencing requires attention to the particular (often-shifting) assemblages of people, institutions, machines and materials that are involved in any particular project. We may therefore develop Fortun's (1999) analysis of the temporalities of genomics. In that, he drew a connection between speed and other factors such as concentration, scale, capital intensivity, and the organisation of labour and space that accelerate the speed of sequencing as well as driving the development and intensification of particular organisational forms such as large-scale sequencing centres.

When one considers sequencing activities more thickly, we may observe different drivers of temporality. Rather than the ever-enhanced speed driven in the thin parts of sequencing by the factors Fortun identifies, alternative priorities may be exhibited. Different organisation of projects and different temporal regimes may be apparent depending on whether we interpret sequencing in a thick or thin manner. In the pig project at the Sanger Institute, the speed of sequencing was halved due to issues of scale and some institutional opposition to the project. This was viewed by many in the pig genome research community as beneficial, as processing and analysis had become – according to SGSC Technical Committee member Craig Beattie – the "rate-limiting step." The speed of production of sequence data meant that they "were overwhelming the information pipeline."[37] So a reduction in speed of production was not a problem. This was, still, fast science, although it was not necessarily so at all stages of the sequencing described in the thick sense. By attending to the thicker understanding of sequencing, one is able to grasp the institutional, collaborative, translational and infrastructural contexts more fully.

In addition, the particularities of the sequencing work in a given community are defined in a sharper and more finely-grained manner, enabling one to identify the conditions that guided particular decisions and actions. In so doing, one can make comparisons between particular objects of study with the aim of defining more precisely how, and to what extent, the conclusions drawn from one may be applicable to the other. To provide one example of this, we may consider two potential objects of study for historians, philosophers and sociologists of science: pig genome research and human genome research. If we were to conduct research based on a thin interpretation of sequencing, both of these objects of study look much the same. The work forming the focus of the thin perspective on sequencing was conducted by specialist teams at large-scale, highly-automated, high-throughput sequencing centres (one of them, the Sanger Institute, participated in both human and pig genome sequencing). One might therefore expect that findings concerning one project will likely be transferable to the other; to re-quote Alan Archibald "to produce a humanised pig genome." Yet based on a thicker study of sequencing, we not only de-humanise the pig genome (research) but genome research altogether. We reveal important differences in library construction, the continuing and leading role of the pig genetics community in the sequencing work (as against the marginalisation of medical geneticists in the human genome project), the rationale for the production of sequence data and the use of allied annotation. The thick perspective also leads us to different characterisations of the projects in terms of scale and velocity.

In this paper I do not claim to establish what sequencing or genomics is, nor to base any of the claims that I do make on a supposed representativeness or significance of pig genome sequencing. I would suggest, however, that the characterisation of sequencing and genomics in much of the scholarly literature is – understandably – dominated by human genome sequencing, and in particular, the efforts that fall under the narrative umbrella of 'The Human Genome Project'. In human genome sequencing, the kinds of work and objects foregrounded by a thin account of

[37] Craig Beattie, Skype interview with author, 23rd March 2017.

sequencing appear to be central, the object of the competition and race between the 'private' and the 'public' initiatives, the area of the work most associated with charismatic and forceful individuals such as John Sulston and Craig Venter, who themselves have helped shape the narratives dominating journalistic and activist discourse (e.g. Sulston and Ferry, 2002; Venter, 2008; see Hilgartner, 2017, chapter 7, for an acute dissection of the narratives; see also a discussion of the "narrative gap" in accounts of the Human Genome Project in Bartlett, 2008, pp. 124-125).

It is precisely those prominent aspects of human genome sequencing that have been associated with scale, automation, speed and many other properties attributed to genomics. Due to the level of funding and the political stakes involved, the imperative to produce sequence data as quickly as possible was more acute for this project than any other sequencing initiative. As the objects and processes highlighted by a thin interpretation of sequencing are proximate to the immediate production of sequence data – the traces transmitted to computers from the bases – it is the stage encompassing these that has gained prominence. To a lesser extent, assembly garnered attention insofar as it was the draft products of this that were announced at the White House in June 2000. Thus, the human genome project was primarily understood in a thin way, and this is consequently how sequencing has become understood.

Finally, I want to emphasise the iterative and recursive nature of sequencing. Sequencing is the production of a tool as well as a dataset. The products of sequencing are intended to be used in scientific investigation for the production of knowledge claims, but also to further improve tools that can be used in investigation and intervention. It is in this sense that further investigation of the development of sequences and sequencing practices towards their intended use and re-use as tools for research and intervention can potentially be fruitful in improving understanding of translational research processes. A thick perspective enables us to open up those processes.

Acknowledgements

For their comments on drafts at various stages, I would like to thank Miguel García-Sancho, Giuditta Parolini and Mark Wong of the TRANSGENE group at the University of Edinburgh, participants in the Perspectives on Genetics and Genomics group, Dominic Berry, Jane Calvert and Doug Lowe. I am additionally grateful for the constructive and helpful comments of the anonymous reviewers. I would like to thank all those who have participated in oral history interviews as part of my research into pig genomics, in particular Alan Archibald of the Roslin Institute and Lawrence Schook of the University of Illinois at Urbana-Champaign who have been generous with their time and allowed me to examine their personal papers. The research for this paper was conducted through the 'TRANSGENE: Medical translation in the history of modern genomics' project which is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 678757. This European research funding is deeply appreciated.

References

Anderson, S. I., Lopez-Corrales, N. L., Gorick, B., & Archibald, A. L. (2000). A large-fragment porcine genomic library resource in a BAC vector. *Mammalian Genome*, *11*, 811–814.

Archibald, A. L., Haley, C. S., Brown, J. F., Couperwhite, S., McQueen, H. A., Nicholson, D., Coppieters, W., Van De Weghe, A., Stratil, A., Winterø, A. K., Fredholm, M., Larsen, N. J., Nielsen, V. H., Milan, D., Woloszyn, N., Robic, A., Dalens, M., Riquet, J., Gellin, J., Caritez, J.-C., Burgaud, G., Ollivier, L., Bidanel, J.-P., Vaiman, M., Renard, C., Geldermann, H., Davoli, R., Ruyter, D., Verstege, E. J. M., Groenen, M. A. M., Davies, W., Høyheim, B., Keiserud, A., Andersson, L., Ellegren, H., Johansson, M., Marklund, L., Miller, J. R., Anderson Dear, D. V., Signer, E., Jeffreys, A. J., Moran, C., Le Tissier, P., Muladno, Rothschild, M. F., Tuggle, C. K., Vaske, D., Helm, J., Liu, H.-C., Rahman, A., Yu, T.-P., Larson, R. G., & Schmitz, C. B. (1995). The PiGMaP consortium linkage map of the pig (*Sus scrofa*). *Mammalian Genome, 6*, 157–175.

Archibald, A. L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M. A., Harlizius, B., Lee, K.-T., Milan, D., Rogers, J., & Rothschild, M. F. (2010). Pig genome sequence-analysis and publication strategy. *BMC genomics, 11,* 1.

Barnes, B., & Dupré, J. (2008). *Genomes and What to Make of Them*. Chicago, IL: The University of Chicago Press.

Bartlett, A. (2008). Accomplishing Sequencing the Human Genome. Unpublished PhD thesis. Cardiff University. Available online at: http://orca.cf.ac.uk/54499/1/U584600.pdf - last accessed 12.05.2017.

Bostanci, A. (2004). Sequencing Human Genomes. In: J.-P. Gaudillière & H.-J. Rheinberger (Eds.) *From Molecular Genetics to Genomics: The mapping cultures of twentieth-century genetics*, pp. 158–179. Abingdon, UK: Routledge.

Bostanci, A. (2006). Two drafts, one genome? Human diversity and human genome research. *Science as Culture*, *15*, 183–198.

Brown, T. A. (2006). *Genomes 3*. New York and London: Garland Science Publishing.

de Chadarevian, S. (2004). Mapping the worm's genome. Tools, networks, patronage. In: J.-P. Gaudillière & H.-J. Rheinberger (Eds.) *From Molecular Genetics to Genomics: The mapping cultures of twentieth-century genetics*, pp. 95–110. Abingdon, UK: Routledge.

Chakravarti, A. (1996). Genetic and Physical Map Integration. In: Abstracts: Swine Chromosome 7 Workshop. *Animal Biotechnology*, *7*, 81–98.

Chow-White, P.A., & García-Sancho, M. (2012). Bidirectional Shaping and Spaces of Convergence: Interactions between Biology and Computing from the First DNA Sequencers to Global Genome Databases. *Science, Technology, & Human Values*, *37*, 124–164.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, *300*, 286–290.

Davis B. D. and colleagues (1990). The Human Genome and Other Initiatives. *Science, 249,* 342–343.

Dawson, H. D., Loveland, J. E., Pascal, G., Gilbert, J. G. R., Uenishi, H., Mann, K. M., Sang, Y., Zhang, J., Carvalho-Silva, D., Hunt, T., Hardy, M., Hu, Z., Zhao, S.-H., Anselmo, A., Shinkai, H., Chen, C., Badaoui, B., Berman, D., Amid, C., Kay, M., Lloyd, D., Snow, C., Morozumi, T., Cheng, R. P.-Y., Bystrom, M.,

Kapetanovic, R., Schwartz, J. C., Kataria, R., Astley, M., Fritz, E., Steward, C., Thomas, M., Wilming, L., Toki, D., Archibald, A. L., Bed'Hom, B., Beraldi, D., Huang, T.-H., Ait-Ali, T., Blecha, F., Botti, S., Freeman, T. C., Giuffra, E., Hume, D. A., Lunney, J. K., Murtaugh, M. P., Reecy, J. M., Harrow, J. L., Rogel-Gaillard, C., & Tuggle, C. K. (2013). Structural and functional annotation of the porcine immunome. *BMC Genomics*, *14*, 332.

Fortun, M. (1999). Projecting Speed Genomics. In: M. Fortun & E. Mendelsohn (Eds.) *The Practices of Human Genetics*, pp. 25–48. Dordrecht, Germany: Springer.

Galison, P., & Hevly, B. (Eds.) (1992). *Big Science: The Growth of Large-Scale Research*. Stanford, CA: Stanford University Press.

García-Sancho, M. (2012). *Biology, Computing and the History of Molecular Sequencing: From Proteins to DNA, 1945-2000*. Basingstoke, UK: Palgrave Macmillan.

García-Sancho, M. (2016). The proactive historian: Methodological opportunities presented by the new archives documenting genomics. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *55*, 70–82.

Gaudillière, J.-P., & Rheinberger, H.-J. (Eds.) (2004). *From Molecular Genetics to Genomics: The mapping cultures of twentieth-century genetics*. Abingdon, UK: Routledge.

Glasner, P. (2002). Beyond the genome: Reconstituting the new genetics. *New Genetics and Society*, *21*, 267–277.

Green, E. D., Guyer, M. S., & National Human Genome Research Institute (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, *470*, 204–213.

Groenen, M. A. (2016). A decade of pig genome sequencing: a window on pig domestication and evolution. *Genetics Selection Evolution*, *48*, 23.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. Li, S., Larkin, D. M., Kim, H., Frantz, L. A. F., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C. W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S.-H., Chow, W., Clark, R. C., Clee, C., Crooijmans, R. P. M. A., Dawson, H. D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z.-Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G. J., Fowler, K. E., Fredholm, M., Fritz, E., Gilbert, J. G. R., Giuffra, E., Gorodkin, J., Griffin, D. K., Harrow, J. L., Hayward, A., Howe, K., Hu, Z.-L., Humphray, S. J., Hunt, T., Hornshøj, H., Jeon, J.-T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J., Kim, J.-H., Kim, K.-W., Kim, T.-H., Larson, G., Lee, K., Lee, K.-T., Leggett, R., Lewin, H. A., Li, Y., Liu, W., Loveland, J. E., Lu, Y., Lunney, J. K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M. P., Narayan, J., Nguyen, D. T., Ni, P., Oh, S.-J., Onteru, S., Panitz, F., Park, E.-W., Park, H.-S., Pascal, G., Paudel, Y., Perez-Enciso, M., Ramirez-Gonzalez, R., Reecy, J. M., Rodriguez-Zas, S., Rohrer, G. A., Rund, L., Sang, Y., Schachtschneider, K., Schraiber, J. G., Schwartz, J., Scobie, L., Scott, C., Searle, S., Servin, B., Southey, B. R., Sperber, G., Stadler, P., Sweedler, J. V., Tafer, H., Thomsen, B., Wali, R., Wang, J., Wang, J., White, S., Xu, X., Yerle, M., Zhang, G., Zhang, J., Zhang, J., Zhao, S., Rogers, J., Churcher, C., & Schook, L. B. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature, 491,* 393–398.

Harrow, J. L., Steward, C. A., Frankish, A., Gilbert, J. G., Gonzalez, J. M., Loveland, J. E., Mudge, J., Sheppard, D., Thomas, M., Trevanion, S., & Wilming, L. G. (2014). The Vertebrate Genome Annotation browser: 10 years on. *Nucleic Acids Research, 42,* D771–D779.

Hilgartner, S. (2013). Constituting Large-Scale Biology: Building a Regime of Governance in the Early Years of the Human Genome Project. *BioSocieties, 8,* 397–416.

Hilgartner, S. (2017). *Reordering Life: Knowledge and Control in the Genomics Revolution*. Cambridge, MA: The MIT Press.

Hogan, A. J. (2014). The 'Morbid Anatomy' of the Human Genome: Tracing the Observational and Representational Approaches of Postwar Genetics and Biomedicine. *Medical History*, *58*, 315–336.

Humphray, S. J., Scott, C. E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A., Patel, M., Sehra, H., Yang, F., Rogatcheva, M. B., Milan, D., Chardon, P., Rohrer, G., Nonneman, D., de Jong, P., Meyers, S. N., Archibald, A., Beever, J. E., Schook, L. B., & Rogers, J. (2007). A high utility integrated map of the pig genome. *Genome Biology*, *8,* R139.

Kuzmuk, K., & Schook, L. (2011). Pigs as a Model for Biomedical Sciences. In: M. Rothschild & A. Ruvinsky (Eds.) *The Genetics of the Pig, Second Edition*, pp 426–444. Cambridge and Oxford, UK: CABI.

Lenoir, T. (1999). Shaping Biomedicine as an Information Science. In M. E. Bowden, T. B. Hahn & R. V. Williams (Eds.) *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (pp. 27-45). Medford, NJ: Information Today, Inc.

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. Chicago, IL: The University of Chicago Press.

McKusick, V. A. (1991). Current trends in mapping human genes. *The FASEB Journal*, *5*, 12–20.

McKusick, V. A. (1997). Mapping the Human Genome: Retrospective, Perspective and Prospective. *Proceedings of the American Philosophical Society, 141*, 417–424.

McKusick, V. A., & Ruddle, F. H. (1987). Toward a Complete Map of the Human Genome. *Genomics, 1*, 103–106.

Meyers, S. N., Rogatcheva, M. B., Larkin, D. M., Yerle, M., Milan, D., Hawken, R. J., Schook, L. B., & Beever, J. E. (2005). Piggy-BACing the human genome II. A high-resolution, physically anchored, comparative map of the porcine autosomes. *Genomics*, *86*, 739–752.

Nadim, T. (2016). Data Labours: How the Sequence Databases GenBank and EMBL-Bank Make Data. *Science as Culture*, *25*, 496–519.

Onaga, L. A. (2014). Ray Wu as Fifth Business: Deconstructing collective memory in the history of DNA sequencing. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *46*, 1–14.

Osoegawa, K., Woon, P. Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J. J., & de Jong, P. J. (1998). An Improved Approach for Construction of Bacterial Artificial Chromosome Libraries. *Genomics*, *52*, 1–8.

Parolini, G. (2018). Building Human and Industrial Capacity in European Biotechnology: The Yeast Genome Sequencing Project (1989–1996). Technische Universität Berlin preprint. http://dx.doi.org/10.14279/depositonce-6693 (Accessed 11 April 2018).

Pool, R., & Waddell, K. (2002). *Exploring Horizons for Domestic Animal Genomics: Workshop Summary*. Washington, D.C.: National Academy Press.

Renard, C., Hart, E., Sehra, H., Beasley, H., Coggill, P., Howe, K., Harrow, J., Gilbert, J., Sims, S., Rogers, J., Ando, A., Shigenari, A., Shiina, T., Inoko, H., Chardon, P., & Beck, S. (2006). The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics, 88,* 96–110.

Rheinberger, H.-J., & Gaudillière, J.-P. (Eds.) (2004). *Classical Genetic Research and its Legacy: The mapping cultures of twentieth-century genetics*. London and New York: Routledge.

Rogel-Gaillard, C., Bourgeaux, N., Billault, A., Vaiman, M., & Chardon, P. (1999). Construction of a swine BAC library: application to the characterization and mapping of porcine type C endoviral elements. *Cytogenetics and Cell Genetics*, *85,* 205–211.

Rohrer, G. A., Alexander, L. J., Hu, Z. Smith, T. P. L., Keele, J. W., & Beattie, C. W. (1996). A Comprehensive Map of the Porcine Genome. *Genome Research*, *6*, 371–391.

Rohrer, G., Beever, J. E., Rothschild, M. F., Schook,, L., Gibbs, R., & Weinstock, G. (2002). Porcine Sequencing White Paper: Porcine Genomic Sequencing Initiative. Available online at: https://www.genome.gov/pages/research/sequencing/seqproposals/porcineseq021203.pdf Last accessed 02.11.2017

Schook, L., Beattie, C., Beever, J., Donovan, S., Jamison, R., Zuckermann, F., Niemi, S., Rothschild, M., Rutherford, M., & Smith, D. (2005a). Swine in biomedical research: creating the building blocks of animal models. *Animal biotechnology, 16,* 183–190.

Schook, L. B., Beever, J. E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohrer, G., & Eversole, K. (2005b). Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comparative Functional Genomics*, *6*, 251–255.

Stein, L. (2001) Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2, 493–503.

Stevens, H. (2011). On the means of bio-production: Bioinformatics and how to make knowledge in a high-throughput genomics laboratory. *BioSocieties*, *6,* 217–242.

Stevens, H. (2013). *Life Out of Sequence: A Data-Driven History of Bioinformatics*. Chicago, IL: The University of Chicago Press.

Strasser, B. J. (2011). The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine. *Isis*, *102*, 60–96.

Sulston, J., & Ferry, G. (2002). *The Common Thread: A Story of Science, Politics, Ethics and the Human Genome*. London, UK: Bantam Press.

Venter, J. C. (2008). *A Life Decoded: My Genome, My Life*. London, UK: Penguin Books.

Vermeulen, N. (2016). Big Biology: Supersizing Science During the Emergence of the 21st Century. *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin*, *24,* 195–223.

Wade, N. (2001). *Life Script: How the Human Genome Discoveries Will Transform Medicine and Enhance Your Health*. New York: Simon & Schuster.

Wellcome Trust (2005). *Strategic Plan 2005-2010: Making a Difference*. London, UK: Wellcome Trust.

Wellcome Trust (2010). *Strategic Plan 2010-2020: Extraordinary Opportunities*. London, UK: Wellcome Trust.

Yerle, M., Lahbib-Mansais, Y., Mellink, C., Goureau, A., Pinton, P., Echard, G., Gellin, J., Zijlstra, C., De Haan, N., Bosma, A. A., Chowdhary, B., Gu, F., Gustavsson, I., Thomsen, P.D., Christensen, K., Rettenberger, G., Hameister, H., Schmitz, A., Chaput, B., & Frelat, G. (1995). The PiGMaP consortium cytogenetic map of the domestic pig (*Sus scrofa domestica*). *Mammalian Genome*, *6*, 176–186.